fl

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Marija Đurđević

# Topološki pristopi k analizi bioloških podatkov

magistrsko delo

študijski program druge stopnje
računalništvo in informatika

Mentor: prof. dr. Nežka Mramor Kosta
Somentor: prof. dr. Blaž Zupan

Ljubljana, 2016

*Magistrski študij je bil pot, ki mi bo za vedno ostala v lepem spominu. Neprecenljiva je bila podpora vseh, ki ste me na tej poti spremljali in spodbujali.*

*Najprej se moram iz srca zahvaliti cenjeni mentorici, prof. dr. Neži Mramor Kosta, za naklonjenost, skrb in spodbudo tekom študija, ter pri izdelavi magistrske naloge. Hvala za vso vašo pomoč in sodelovanje, predvsem pa za razumevanje in potrpežljivost.*

*Prav tako se iskreno zahvaljujem somentorju in izjemnemu predavatelju, prof. dr. Blažu Zupanu, za strokovno pomoč in preneseno znanje. Najlepša hvala še dr. Marinki Žitnik za kakovostno podporo in nasvete.*

*Zahvala gre tudi mojemu fantu, ki me je spodbujal pri delu in verjel v moj uspeh.*

*Posebno zahvalo pa namenjam svoji družini za neizčrpno ljubezen in razumevanje na vsakem koraku moje poti.*

Mojoj porodici.

# Kazalo

# Seznam uporabljenih kratic

| kratica | angleško | slovensko |
|---|---|---|
| **ICGC** | The International Cancer Genome Consortium | Mednarodni konzorcij za genske raziskave raka |
| **DLBCL** | Diffuse Large B-cell Lymphoma | |
| **TDA** | Topological Data Analysis | Topološka analiza podatkov |
| **VR** | Vietoris-Rips | Vietoris-Ripsov kompleks |
| **PCA** | Principal Components Analysis | Analiza glavnih komponent |
| **DNA** | Deoxyribonucleic Acid | Deoksiribonukleinska kislina |
| **RNA** | Ribonucleic Acid | Ribonukleinska kislina |
| **BRCA** | Breast Cancer | Rak prsi |
| **OV** | Ovary Cancer | Rak jajčnikov |
| **LUSC** | Lung Squamous Cell Carcinoma | Pljučni ploščatocelični karcinom |
| **LAML** | Acute Myeloid Leukemia | Akutna mieloična levkemija |
| **LOWESS** | Locally Weighted Scatterplot Smoothing | Lokalno utežena regresija |
| **CSR matrix** | Compressed Sparse Row matrix | Redke stisnjene matrike |

# Seznam uporabljenih simbolov

| **simbol** | **opis** |
|---|---|
| $T$ | topološki prostor |
| $k$ | dimenzija |
| $X$ | množica točk |
| $S$ | množica simpleksov |
| $\sigma$ | simpleks |
| $\tau$ | podsimpleks |
| $\alpha$ | abstraktni simpleks |
| $\beta$ | abstraktni podsimpleks |
| $K$ | geometrijski simplicialni kompleks |
| $A$ | abstraktni simplicialni kompleks |
| $M$ | metrični prostor |
| $\varepsilon$ | parameter bližine |
| $\partial$ | rob |
| $v$ | vozlišče |

# Povzetek

Podatki o genski izraženosti rakavega tkiva imajo napovedno vrednost pri napovedovanju bolnikovega kliničnega izida. Na področju rakavih bolezni je pomembno ugotavljanje podkategorije v posamezni kategoriji raka. V magistrski nalogi smo se za reševanje tega problema odločili za implementacijo algoritmov, ki temeljijo na računski topologiji. Cilj naloge je, da z računanjem vztrajne homologije na podatkih o genski izraženosti rakavega tkiva ugotovimo nove podskupine ter poskusimo napovedovati preživetje bolnikov v skupinah. Podatki, ki smo jih analizirali, izhajajo iz podatkovne zbirke mednarodnega konzorcija za genske raziskave raka ICGC. Na omenjenih podatkih smo gradili simplicialne komplekse pri različnih resolucijah z uporabo algoritma Vietoris-Rips. Nato smo računali vztrajno homologijo in izrisovali vztrajne diagrame. Z namenom, da čim bolj natančno ločimo podkategorije raka, smo razvili metodo za računanje intervala zaupanja na vztrajnih diagramih. Na ta način smo uspešno odkrili nekaj novih podskupin ter napovedali klinični izid bolnikov. Uspeh metod smo ovrednotili na podatkih z več različnimi tipi raka ter rezultate uspešno primerjali z drugimi metodami nenadzorovanega učenja.

**Ključne besede:** topologija, topološka analiza podatkov, simplicialni kompleks, Vietoris-Rips, rak, klasifikacijske metode, krivulja preživetja.

# Abstract

**Title:** Topological Approach to Analyses of Omics Data

Genes expression is often a good indicator for prediction of patient's clinical results. In diseases such as cancer is inevitable to identify subcategories of phenotype. The goal of the thesis is to use persistent homology on cancer tissue gene expression to identify new subgroups and try to predict the survival of patients in corresponding groups. The data was obtained from the International Consortium for Cancer Research. Simplicial complexes were built for different resolutions using Vietoris-Rips algorithm. We counted the persistent homology and draw persistent diagrams. A method for calculating confidence interval on persistent diagrams was developed to precisely divide cancer subcategories. This method gave us promising results by discovering new subcategories and was accurate in prediction of patient clinical results. Results were obtained on data of different cancer types and compared with several unsupervised learning methods.

**Keywords:** topology, topological data analysis, simplicial complex, Vietoris-Rips, cancer, classification methods, survival curves.

# Poglavje 1

# Uvod

Navkljub razvoju medicinske znanosti, ki je v nekaj zadnjih desetletjih naredila velik napredek v razumevanju bolezni raka, je smrtnost zaradi te bolezni še vedno visoka [24, 10]. Ena od ključnih stvari v napovedovanju bolnikovega kliničnega izida je odkrivanje podtipov v posameznem tipu raka. Razvrstitev raka samo na podlagi morfoloških karakteristik se je izkazala kot nezanesljiva.

Pojav genomskih tehnologij, kamor spadajo tudi mikromreže, je odprl novo poglavje v analizi te krute bolezni. Mikromreže hkratno spremljajo izražanja več tisoč genov [23]. Pridobljeni podatki so zelo kompleksni za analizo, ker vsebujejo veliko šuma. Do sedaj uporabljane klasifikacijske metode so se izkazale kot dokaj uspešne, vendar pri nekaterih vrstah raka te metode ne ločijo dovolj dobro posameznih podtipov [15]. Na primer, klinični rezultati tumorjev, ki so razvrščeni kot rak limfe (ang. *Diffuse Large B-cell Lymphoma DLBCL*), so zelo spremenljivi, kar kaže, da obstaja več podtipov tega tipa raka [3]. Želimo namreč poiskati skupine, ki so čim bolj ločene druga od druge, hkrati pa v njih razkriti podskupine, za katere bomo napovedovali klinične izide.

V magistrski nalogi smo k reševanju tega problema pristopili z algoritmi, ki temeljijo na algebrski topologiji. Podatke o genski izraženosti smo pridobili iz podatkovne zbirke Mednarodnega konzorcija za genske raziskave raka (ang. *The International Cancer Genome Consortium ICGC*). Podatkovna zbirka

je uvrščena v trenutno najbolj aktualne in do sedaj še ni bila analizirana z uporabo topoloških metod.

Topološka analiza podatkov se je kot relativno nov pristop k analizi podatkov v zadnjem času izkazala za zelo uspešno v analizi kompleksnih bioloških podatkov [2]. Najbolj znana je uporaba metode Mapper v delu M. Nicolau in ostalih [17] pri analizi napovedovanja bolezni raka prsi. Alpha kompleksov in Delaunayeve triangulacije uspešno določajo mesta interakcije med beljakovinami [26]. L. Seemann in soavtorji [18] z uporabo vztrajne homologije napovedujejo tipe raka na podlagi genske ekspresije majhne množice genov. L. Li in soavtorji [13] topološkom analizom podobnosti bolnikov identificirajo podtipe diabetesa tipa dve.

Magistrska naloga temelji na kombinaciji aktualnih algoritmov računske topologije ter drugih algoritmov za vrednotenje rezultatov nenadzorovanega učenja. V začetni fazi za izgradnjo simplicialnih kompleksov uporabljamo algoritem Vietoris-Rips. Pri različnih resolucijah nato računamo vztrajno homologijo, katera uspešno obravnava šum v podatkih in razvrsti podatke v skupine. Uporabom metode za računanje intervala zaupanja na vztrajnih diagramih zanemarimo topološke značilnosti kratke življenjske dobe (šum). Simplekse dimenzije nič, ki gradijo topološke značilnosti z večjim časom preživetja, predlagamo kot potencialne gruče bolnikov. Uspešnost rezultata preverjamo metodom „Silhouette“ ter napovedujemo preživetje bolnikov računanjem Kaplan-Meierjevih krivulj.

Naštete metode so implementirane v programskem jeziku `Python`. Za obdelavo matrik smo uporabili knjižnice `numpy` in `scipy`, za vrednotenje gruč knjižnico `scikit-learn` ter knjižnico `lifelines` za napovedovanje preživetja bolnikov. Slike so izrisovane v spletnem orodju `Goegebra`.

V nalogi najprej na kratko predstavimo topološke pristope k analizi bioloških podatkov. Nato opišemo gradnjo simplicialnega kompleksa z uporabo algoritma Vietoris-Rips ter računanja vztrajne homologije. V nadaljevanju predstavimo metodo za analizo vztrajnih diagramov, ki smo jih uporabljali v sklopu magistrske naloge. V zadnjem poglavju poročamo o rezultatih in

uspešnosti napovedovanja kliničnega izida bolnikov ter primerjavi z drugimi metodami nenadzorovanega učenja. Nalogo zaključimo s kratkim povzetkom ugotovitev.

# Poglavje 2

# Topološki pristopi v analizi podatkov

Topološki pristop k analizi podatkov je razmeroma novo področje, ki se razteza čez več disciplin, vključno s topologijo, računsko geometrijo, statistiko in strojnim učenjem. Zasnovano je na naraščajoči množici učinkovitih algoritmov, ki zagotavljajo vpogled v ,,obliko“ podatkov. Topološka analiza podatkov (ang. *Topological Data Analysis TDA*) se je izkazala kot zelo uspešna v odkrivanju informacij pri visoko dimenzionalnih podatkih [5]. V primerjavi z drugimi metodami (kot je recimo PCA), ki poskušajo zmanjšati dimenzionalnost podatkov, TDA omogoča analizo v polni dimenziji ter brez izgube pomembnih informacij.

## 2.1 Topološka analiza podatkov

Ključne prednosti uporabe topoloških metod so brezkoordinatni opis podatkov v prostoru, robustnost v prisotnosti šuma in invariantnost glede na vrsto transformacij. Homološka analiza podatkov omogoča odkrivanje topoloških značilnosti, ki temeljijo na povezanosti. To so cikli, luknje, tuneli in druge visoko dimenzionalne oblike, ki jih ni mogoče zaznati s tradicionalnimi metodami, kot je gručenje. Sledenje življenjski dobi topoloških značilnosti za-

gotavlja, da izberemo signifikantne ,,oblike“, medtem ko zanemarimo šum.

Zaradi številnih prednosti je topološka analiza podatkov našla široko uporabo v različnih področjih. Topološki modeli, kot so Alpha oblike, Delaunayeva triangulacija in algoritmi za njihovo analizo, se že nekaj let uspešno uporabljajo tudi za reševanje problemov s področja bioinformatike. Izkazali so se za zelo uspešne v določanju interakcije med beljakovinami. Navedene metode niso samo robustne tehnike za rekonstrukcijo interakcij beljakovin, temveč se uporabljajo tudi za opisovanje interakcijskih točk [26].

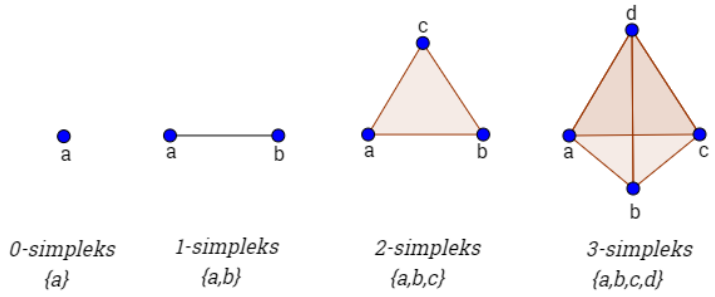
V raziskavi napredovanja bolezni raka je bila uspešno uporabljena topološka metoda Mapper za analizo in vizualizacijo podatkovnih nizov visoke dimenzije. Ti so pokazali, da topološka analiza podatkov omogoča pogled na podatke, ki je bolj enostaven za razumevanje, ter da so topološki opisi podatkov v šibki povezavi z biološkimi opisi [17].

## 2.2 Algoritem Vietoris–Rips

Podatke predstavimo kot točke prostora $\mathbb{R}^n$, kjer bližnje točke, ki ustrezajo podobnim podatkom, med seboj povežemo v strukturo imenovano simplicialni kompleks. Topološka analiza opiše ,,obliko“ tega prostora z iskanjem povezanih komponent, lukenj, tunelov in drugih večdimenzionalnih značilnosti v prostoru. V nadaljevanju navajamo definicijo Edelsbrunnerja in Harerja [5]:

**Definicija 2.1** *Konveksno ovojnico $k$+1 afino neodvisnih točk $S = \{v_0,v_1,...,v_k\}$ imenujemo* ***(geometrijski) k-simpleks****, kjer $k$ označuje dimenzijo simpleksa. Pri tem so točke $v_0, v_1, \ldots, v_k \in \mathbb{R}^n$ afino neodvisne, če ne obstajajo taka neničelna števila $\alpha_0, \ldots, \alpha_k$ z vsoto $\sum_{i=0}^{k} \alpha_i = 0$, da je $\sum_{i=0}^{k} \alpha_k v_k = 0$.*

Iz definicije je razvidno, da so simpleksi dimenzije 0 točke, dimenzije 1 povezave, dimenzije 2 trikotniki in tako naprej (slika 2.1).




Slika 2.1: Orientirani $k$-simpleksi.

**Definicija 2.2** *Končna družina $K$ simpleksov je (geometrijski) simplicialni kompleks, če zadošča spodnjim zahtevam:*

1. $\sigma \in K, \tau \leq \sigma \implies \tau \in K,$

2. $\sigma_1, \sigma_2 \in K \implies \sigma_1 \cap \sigma_2 \in K.$

Na primer, če trikotnik $\triangle abc$ pripada množici $K$, potem morajo biti vse povezave $ab$, $bc$, $ca$ in vozlišča $a$, $b$ in $c$ iz $K$. Druga lastnost pa nam pove, da sta dva $k$-simpleksa ločena, ali pa je njihovo presečišče nižje dimenzionalni simpleks, ki pripada simplicialnemu kompleksu $K$.

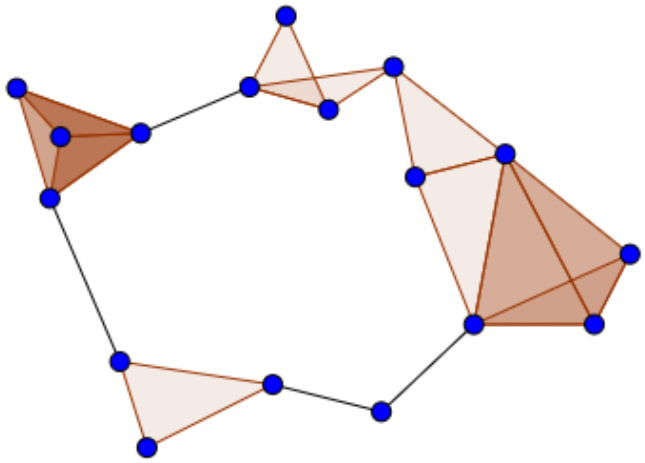
Simplicialni kompleks je topološki prostor, zgrajen z ,,lepljenjem" točk, daljic, trikotnikov in simpleksov višje dimenzije. Primer simplicialnega kompleksa je prikazan na sliki 2.2.

Geometrijski simplicialni kompleks je geometrijska realizacija abstraktnega simplicialnega kompleksa.

**Trditev 2.1** *Abstraktni simplicialni kompleks dimezije $k$ je mogoče geometrijsko realizirati v prostoru $\mathbb{R}^{2k+1}$.*

Edelsbrunner in Harer [5] definirata abstraktni simplicialni kompleks na naslednji način:

**Definicija 2.3** ***Abstraktni simplicialni kompleks*** *je družina takšnih končnih množic $A$, da če je $\alpha \in A$ in $\beta \subseteq \alpha \implies \beta \in A$. Množice, ki pripadajo $A$, imenujemo abstraktni simpleksi.*

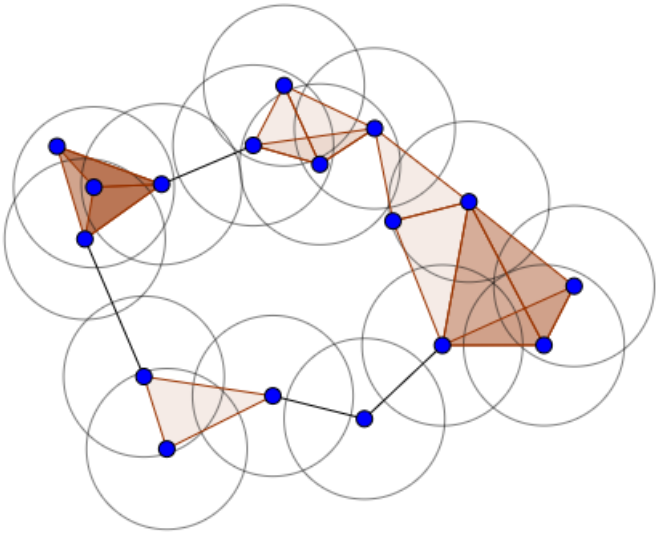


Slika 2.2: Simplicialni kompleks. Modre točke predstavljajo simplekse dimenzije 0, črne povezave so simpleksi dimenzije 1. S svetlo rdečo in temno rdečo barvo so prikazani simpleksi dimenzije 2 in 3, respektivno.

V magistrskem delu smo za izgradnjo simplicialnega kompleksa z oglišči v dani množici točk uporabljali tako imenovani Vietoris-Ripsov kompleks. Le-ta je primer abstraktnega simplicialnega kompleksa.

V Vietoris-Ripsovem pristopu na začetku določimo parameter $\varepsilon > 0$, ki je **parameter bližine** (ang. *proximity parameter*). Nato simplicialni kompleks $K_\varepsilon$ gradimo na naslednji način. V $K_\varepsilon$ je vsaka množica $k+1$ točk metričnega prostora $M$ $k$-simplex, če je razdalja med točkami manj kot $\varepsilon$. Tako 0-simplekse predstavljajo same točke. 1-simpleksi so vse povezave med točkami, pri katerih je razdalja manjša kot $\varepsilon$. 2-simpleksi tvorijo vsi trikotniki, kadar so pari točk v trikotniku med seboj oddaljeni za manj kot $\varepsilon$. 3-simpleksi (tetraedri) nastanejo, če so štiri točke znotraj $\varepsilon$ paroma druga do

druge (slika 2.3).



Slika 2.3: Simplicialni kompleks, zgrajen z uporabo Vietoris-Ripsovega algoritma. Okoli vsake točke je narisan krog radija $\varepsilon$. Če se dva kroga sekata, pomeni, da so pripadajoče točke paroma oddaljene za manj kot $\varepsilon$.

## 2.3 Vztrajna homologija

Homologija je način za odkrivanje $k$-dimenzionalnih lukenj v simplicialnem kompleksu. Predstavljajmo si, da imamo simplicialni kompleks $K$ dimenzije $k$. Ustvarimo lahko abstraktni vektorski prostor $C_k$, z bazo, ki jih sestavljajo množice $k$-simpleksov iz $K$ tako, da je dimenzija $C_k$ enaka številu $k$-simpleksov. Elementi tega vektorskega prostora se imenujejo $k$-verige. $K$-veriga je po definiciji linearna kombinacija:

$$c = \sum_{i=1}^{k} a_i \sigma_i, \tag{2.1}$$

kjer je $a_i$ koeficient, ki bo v tem delu iz obsega $\mathbb{Z}_2$ (torej 0 ali 1). Za računanje homologije je nujno definirati še rob $k$-simpleksa [28].

Rob $k$-simpleksa $\sigma = [v_0,v_1, \ldots,v_k]$ je unija $(k-1)$-simpleksov $\tau \subset \sigma$. Algebrajsko opišemo operacijo, ki simpleksom priredii njihov rob, tako, da za vsako naravno število $k$ definiramo linearno preslikavo $\partial_k : C_k \to C_{k-1}$, ki $k$-simpleksu $\sigma$ priredi $(k-1)$-verigo:

$$\partial_k([v_0, ..., v_i, ..., v_k]) = \sum_{i=0}^{k} (-1)^i [v_0, ..., \hat{v_i}, ..., v_k], \tag{2.2}$$

kjer $\hat{v_i}$ označuje, da smo oglišče $v_i$ izpustili.

Na primer:

$$\partial_0([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]. \tag{2.3}$$

Tu $[v_0,v_1,v_2]$ predstavlja trikotnik (simpleks dimenzije 2) in $[v_1,v_2]$, $[v_0,v_2]$, $[v_0,v_1]$ orientirane povezave (simplekse dimenzije 1), ki tvorijo njegov rob.

Vetorski prostori $C_k$ so v posebnem primeru tudi grupe, robne preslikave $\partial_k$ pa so homomorfizmi grup. Poleg grup $C_k$ sta za nas pomembni še dve družini grup:

1. grupa $k$-ciklov $Z_k(K) = \ker \partial_k$: $C_k \to C_{k-1}$,
2. grupa $k$-robov $B_k(K) = \text{im } \partial_{k+1}$: $C_{k+1} \to C_k$.

Za robni homomorfizem velja, da je $\partial_{k-1} \circ \partial_k = 0$, zato je $B_k \subset Z_k$ za vsak $k \geq 1$. Pokažimo, da res velja $\partial_k \circ \partial_{k+1} = 0$:

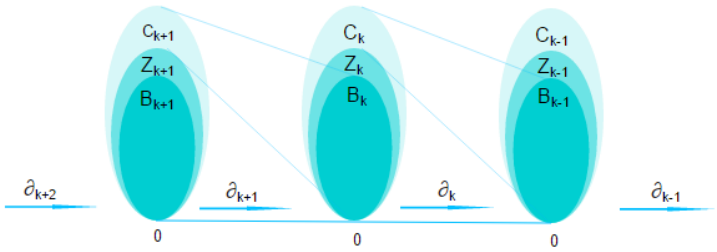
$$\begin{aligned} \partial_{k-1}\partial_k = \partial_{k-1} \sum_{i=0}^{k} (-1)^i [v_0, ..., \hat{v_i}, ..., v_k] = \\ \sum_{j<i}^{k} (-1)^i (-1)^j [v_0, ..., \hat{v_j}, ..., \hat{v_i}, ..., v_k] + \\ \sum_{j>i}^{k} (-1)^i (-1)^{j-1} [v_0, ..., \hat{v_i}, ..., \hat{v_j}, ..., v_k] = 0, \end{aligned} \tag{2.4}$$

saj v vsaki od obeh vsot isti simpleks nastopi z nasprotnim predznakom.

Pokažimo še na primeru trikotnika:

$\partial_1 \circ \partial_2 ([v_0, v_1, v_2]) = \partial_1([v_1, v_2])$ - $\partial_1([v_0, v_2]) + \partial_1([v_0, v_1]) = [v_2]$ - $[v_1]$ - $[v_2] + [v_0] + [v_1]$ - $[v_0] = 0$.

Pri drugem pogoju sledi, da je $B_k$ podprostor prostora $Z_k$. $C_k$ je vektorski prostor vseh $k$-verig v simplicialnem kompleksu $K$, $Z_k$ je podprostor prostora $C_k$, ki sestoji iz $k$-verig, katere so tudi $k$-cikli, in $B_k$ je podprostor prostora $Z_k$, sestavljenega iz $k$-ciklov, ki so tudi $k$-robovi (slika 2.4).




Slika 2.4: Robni homomorfizem (povzeto iz [9]).

$K$-ta homološka grupa simplicialnega kompleksa $K$ (pravzaprav verižnega kompleksa $C(K)$) je kvocientna grupa grupe ciklov $Z_k(K)$ po podgrupi robov $B_k(K)$. Definicije kvocientne grupe tu ne navajamo, saj presega okvire tega magistrskega dela, podroben opis konstrukcije homoloskih grup najdemo v knjigi Edelsbrunnerja in Harerja [5].

$$H_k(K) = \frac{Z_k(K)}{B_k(K)}. \tag{2.5}$$

Element homološke grupe (homološki razred) je torej cel ekvivalenčni razred ciklov in ga lahko predstavimo s poljubnim ciklom iz njegovega ekvivalenčnega razreda. Tako je tudi $H_k$ vektorski prostor, dimenzijo tega prostora pa imenujemo $k$-to Bettijevo število simplicialnega kompleksa $K$: $\beta_k(K) =$ rang $H_k(K)$. Velja:

$$\beta_k(K) = rang Z_k(K) - rang B_k(K). \tag{2.6}$$

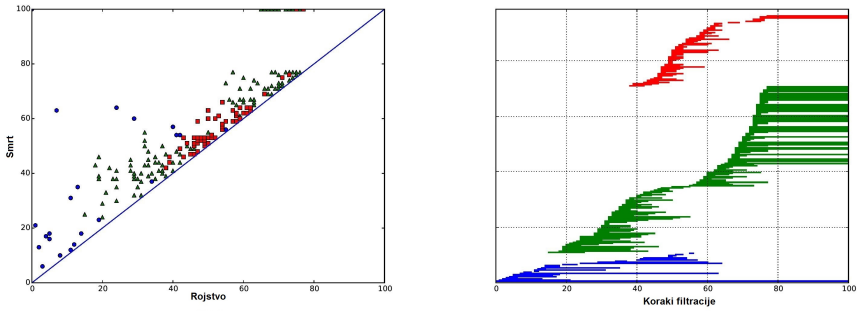
Topološko gledano nam Bettijevo število daje opis oblike simplicialnega kompleksa. Torej, Bettijevo število dimenizije 0, ki je ključnega pomena za našo analizo, nam pove število povezanih komponent. Bettijevo število dimenzije 1 pomeni število tunelov, Bettijevo število 2 število 2-dimenzionalnih votlin in tako dalje. Oblika simplicialnega kompleksa se lahko opiše z zaporedjem Bettijevih števil.

Vztrajna homologija je metoda za izračun topoloških značilnosti prostora v različnih prostorskih resolucijah. Filtracija simplicialnega kompleksa $K$ je zaporedje simplicialnih podkompleksov:

$$\emptyset = K_0 \subset K_1 \subset \cdots \subset K_n = K. \tag{2.7}$$

V našem primeru so $K_i$ simplicialni kompleksi, ki jih dobimo pri različnih resolucijah z Vietoris–Ripsovim algoritmom. Različne nivoje filtracije dosežemo s spreminjanjem praga $\varepsilon$. Zatem sledimo spreminjanju homoloških grup topološkega prostora.

Spreminjanje homoloških grup skozi čas (tekom filtracije) grafično predstavimo na vztrajnih diagramih ali s črtnimi kodami. Za homološki razred z vztrajnostjo $h_j - h_i$ narišemo točko v ravnini s koordinatama $(h_i,h_j)$, kjer $h_i$ imenujemo rojstvo, $h_j$ pa smrt homološkega razreda (slika 2.5). Lahko pa ga predstavimo tudi z uporabo črte dolžine $h_j - h_i$. Topološke razrede, ki preživijo do konca, ponazorimo s parom $(h_i,\infty)$.




Slika 2.5: Vztrajni diagram (levo) in njemu odgovarjajoči graf s črtnimi kodami (desno).

# Poglavje 3

# Metode, orodja in podatki

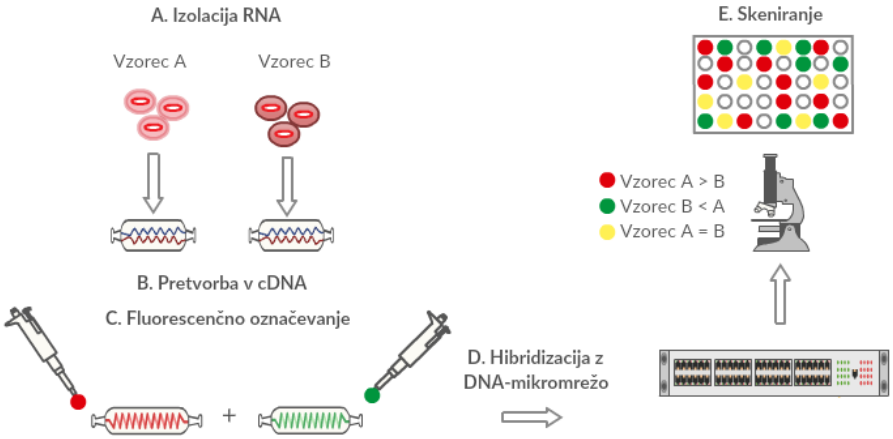
V tem poglavju opišemo uporabljene podatke ter računske metode in orodja, ki smo jih uporabili za gradnjo modelov.

## 3.1 Podatki

Podatki o genski izraženosti so lahko eni izmed ključnih za natančno diagnostično in prognostično oceno raka [12]. Danes obstajajo številne tehnologije, osnovane na genskih mrežah (DNA mreže, mikro mreže, makro mreže ...), ki nam omogočajo merjenje genske izraženosti v tkivu. Genske mreže uporabljamo za identifikacijo sekvence (npr. pri analizi mutacij) ali za diferencialno analizo ekspresije dveh ali več RNA vzorcev. Omogočajo nam močan pristop v primerjavi kompleksnih RNA vzorcev, genski izraženosti normalnega in rakavega tkiva, ter različne razvojne stopnje tkiva. Slika 3.1 prikazuje postopek merjenja genskih izrazov z genskimi mrežami.

Ker z genskimi mrežami sočasno pridobimo na tisoče podatkovnih točk v enem poskusu in ker je ta tehnologija čedalje bolj dostopna in cenejša, podatkovne baze, ki hranijo podatke o genski izraženosti, eksponencialno naraščajo. Glavni izziv predstavlja odkriti znanje iz takih podatkov. V magistrski nalogi smo se izziva lotili s topološko analizo. Podatke smo pridobili iz najbolj aktualne prosto dostopne podatkovne zbirke Mednarodnega konzor-

cija za genske raziskave raka – ICGC[1]. Podatkovna baza je zelo kompleksna in združuje podatke iz različnih virov ter pridobljene z različnimi tehnologijami. Podatki so pogosto nepopolni, še posebej podatki o preživetju bolnikov.




Slika 3.1: Postopek dobivanja podatkov iz genskih mrež. Pri mikromrežah najprej izoliramo mRNA (A), potem pa sledi pretvorba v cDNA (B). Fluorescenčno označimo cDNA (C) in naredimo hibridizacijo z DNA-mikromrežo (D). Nato sledi skeniranje hibridiziranega rastra in interpretacija slike (E).

[1] `https://dcc.icgc.org`

ICGC trenutno vsebuje podatke o 17.867 osebah (od tu naprej bomo uporabljali izraz donator), pridobljenih v 66 različnih projektih iz 16 držav, razvrščenih v 21 primarnih tipov raka. Za potrebe magistrske naloge smo se odločili za analizo podatkov o treh najbolj pogostih tipih raka: o prsnem raku – BRCA, raku na jajčnikih – OV in pljučnem raku – LUSC. Da bi zmanjšali podatkovni nabor ter s tem časovno trajanje analize, smo od vsakega od naštetih tipov raka izbrali po en projekt.

## 3.2 Priprava podatkov

Podatki o genski izraženosti podatkovne baze ICGC so pridobljeni z različnimi tehnologijami, so različnega podatkovnega tipa ter imajo različno strukturo. Od vsakega donatorja je vzetih več vzorcev iz rakavega in zdravega tkiva. Vzorci donatorjev so tudi iz različnih časovnih obdobij. Pri prvih so vzeti v začetni fazi bolezni, potem pa je po nekem času vzorčenje ponovljeno. Pri drugih pa so podatki samo iz začetne ali samo zadnje faze bolezni. Pri številnih pa čas vzorčenja ni znan - znani so samo podatki o genski izraženosti.

Da bi zmanjšali heterogenost podatkov in na ta način izboljšali rezultate, smo podatke najprej ustrezno obdelali, kot je opisano v nadaljevanju in pripravili za analizo. Podatke, razvrščene v različne tabele, smo na začetku združili. Nato smo izločili vzorce donatorjev, ki so samo iz rakavega tkiva ter pridobljeni z isto tehnologijo. V našem primeru, pridobljeni z uporabo Agilent 244K Custom Gene Expression G4502A-07-3 (BRCA, LUSC, OV) platforme za merjenje genske izraženosti. Na ta način smo zmanjšali heterogenost podatkov, ampak še vedno nismo dosegli primerljivosti med podatki. Podatkovni nabor je še zmeraj vseboval časovno različne vzorce istega donatorja.

Da bi dosegli cilj magistrske naloge in z uporabo topoloških metod na podlagi genske ekspresije ugotovili podobnosti med donatorji, smo izbrali samo po en vzorec vsakega donatorja. Odločili smo se, da bo to prvi vzorec,

vzet od donatorja. V primerih, kjer ni bilo dodatnih informacij, smo na podlagi identifikacijske številke vzorca ugotavljali, kateri je prvi vzorec – izbrali smo torej tistega z najmanjšo identifikacijsko številko. Na ta način smo dosegli, da vsaj večina vzorcev pripada začetni fazi bolezni.

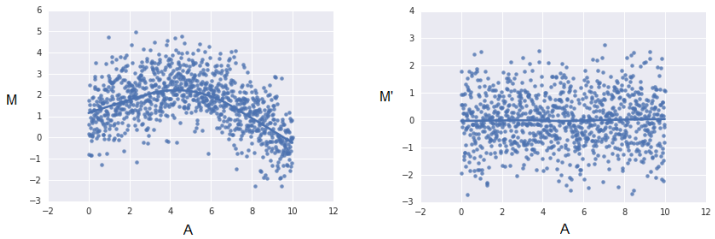
Podatke smo predstavili v obliki matrike za vsak tip raka posebej, kjer vrstice predstavljajo donatorje, stolpci pa gene. Elementi so vrednosti genske izraženosti pri določenem donatorju za dani gen.

Pomembno je tudi poudariti, da so vrednosti genske izraženosti normalizirane z uporabo normalizacije ,,lowess". Tekom branja podatkov iz genske mreže (slika 3.1) pride do številnih napak. Napake nastanejo zaradi kemijskega pojava, kjer barve molekul v neposredni bližini absorbirajo svetlobo druga od druge, s čimer zmanjšajo signal. Najbolj ugoden način za prikaz odvisnosti intenzitete barv je t.i. MA graf, kjer na ordinatni osi prikažemo logaritemsko razmerje (M) intenziteta signalov rdeče (R) in zelene (Z) barve, na abscisni osi pa logaritem njune povprečne intenzitete (A) (slika 3.2).

$$M = \log_2(\frac{R}{Z}), A = \log_2(\frac{R * Z}{2}). \tag{3.1}$$

Lokalno utežnom regresijom na MA grafih popravimo napake (slika 3.2 desno):

$$\log_2(r^*) = \log_2(\frac{R}{Z}) - \text{lowess}(R * Z). \tag{3.2}$$




Slika 3.2: MA graf pred normalizacijo (levo) in po normalizaciji (desno). Modra črta predstavlja ,,lowess" krivuljo izraženosti.

V tabeli 3.1 je podana velikost matrik za vsak tip raka. Za vsako matriko je narejena tudi tabela, ki vsebuje metapodatke o donatorjih. Podatki so shranjeni v formatu `tsv` (ang. *tab-separated values*). Koda za obdelavo in pripravo podatkovne množice je napisana v programskem jeziku `Python`.

| Tip raka | Število donatorjev | Število genov | Platforma |
|---|---|---|---|
| BRCA | 65 | 17.662 | Agilent 244K Custom Gene Expression G4502A-07-3 |
| LUSC | 65 | 11.862 | Agilent 244K Custom Gene Expression G4502A-07-3 |
| OV | 292 | 11.876 | Agilent 244K Custom Gene Expression G4502A-07-3 |

Tabela 3.1: Podatki o donatorjih.

## 3.3 Računanje razdalje

Dodatna prednost topoloških metod je fleksibilnost merjenja razdalje. To pomeni, da nismo vezani na določeno metriko in metrični prostor, kot je recimo pri razvrščanju v skupine z **metodo voditeljev** (ang. *k-means*), ki uporablja evklidsko metriko. Z uporabo kakršne koli druge metrike se pri tem algoritmu lahko zgodi, da algoritem ne skonvergira [21] [19].

V primeru bioloških podatkov se je evklidska metrika izkazala kot ne dovolj ustrezna. Najpogosteje moramo metriko določiti v skladu s samimi podatki. Glede na lastnosti podatkov smo za mero razdalje med dvema točkama $d(x, y)$ izbrali korelacijsko razdaljo med donatorji:

$$d(x,y) = 1 - r_{x,y} = 1 - \frac{\sum_{n=1}^{n}(x_i - x)(y_i - y)}{\sqrt{[\sum_{n=1}^{n}(x_i - \bar{x})^2][\sum_{n=1}^{n}(y_i - \bar{y})^2]}}, \tag{3.3}$$

kjer $r_{x,y}$ predstavlja Pearsonov korelacijski koeficient, ki se najpogosteje uporablja pri analizi podatkov o genski izraženosti. Razlog za to je njegova občutljivost na osamelce ter zaznavanje linearnih povezav v podatkih. Poleg tega Pearsonova korelacija predpostavlja, da so podatki o genski izraženosti normalno porazdeljeni [8] [22].

Pri računanju razdalje smo uporabljali `Pythonovo` knjižnico `numpy`. Razdalje med donatorji smo shranili v obliki matrik razdalje.

## 3.4 Izgradnja Vietoris-Ripsovega simplicialnega kompleksa

Kot je opisano v podpoglavju 2.2, je Vietoris-Ripsov kompleks (VR kompleks) abstraktni simplicialni kompleks, zgrajen na množici $S$, kjer simplekse tvorijo tiste podmnožice množice $S$, ki imajo premer manjši ali enak od nekega izbranega praga $\varepsilon \in \mathbb{R}$.

Obstajata dva pristopa pri gradnji VR kompleksa. Pri prvem pristopu najprej dodamo vse točke topološkega prostora, nato pa na podlagi izbranega premera $\varepsilon$ dodajamo preostale simplekse oz. večdimenzionalne oblike.

Pri drugem pristopu, uporabljenem tudi v magistrski nalogi, dodajamo točke množice $S$ v VR kompleks kot simplekse dimenzije 0 samo kot oglišča stranic, torej v parih z oglišči, oddaljenimi za manj kot $\varepsilon$. Z uporabo tega pristopa smo se izognili problemu, da imamo v filtraciji in računanju vztrajne homologije na začetku veliko število komponent, ki predstavljajo posamezne točke.

Glede na kombinatorično strukturo topoloških modelov imajo algoritmi pogosto veliko časovno in prostorsko kompleksnost. Za izgradnjo VR kompleksa obstajajo številni algoritmi. V članku je opisan algoritem s časovno kompleksnostjo $O(M(n))$ in prostorsko kompleksnostjo reda $n^2$ [27]. Vendar se v našem primeru, kjer so podatki zelo korelirani ter obogateni, ni izkazal za dovolj učinkovitega. Zaradi tega smo problem poskusili rešiti z uporabo algoritma za računanje z redkimi matrikami, opisanega v članku N. Bella

in A. N. Hiranija [1]. Algoritem je bilo potrebno ustrezno prilagoditi naši metriki in podatkovni strukturi.

Pri algoritmu iz matrike razdalj najprej izluščimo podatke o tem, katera vozlišča bodo vključena v 0-skelet VR kompleksa. To predstavimo v obliki matrike $E$, kjer je $E[i,j] = 1$, če sta vozlišči $v_i$ in $v_j$ oddaljeni za manj kot $\varepsilon$ in če je $v_i < v_j$, sicer je 0. Podatki $E$ so shranjeni v podatkovni strukturi za redke matrike csr($E$), dimenzije $3 \times m$, kjer je $m$ število enic v matriki $E$, oziroma število povezav. Prva vrstica matrike csr($E$) vsebuje indekse vrstic druga pa indekse stolpcev matrike $E$, v katerih so vrednosti različne od 0. Tretja vrstica hrani podatke, v našem primeru same enice.

Matrika $E$ ter ustrezna redka matrika csr($E$) za simplicialni kompleks, podan na sliki 3.3:

$$E = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \text{csr}(E) = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Iz matrike $E$ lahko sedaj ustvarimo $k$-skelet VR kompleksa v obliki matrike $F_p$ dimenzije $m \times n$, kjer je $n$ število vozlišč, $m$ pa število $p$-simpleksov v VR kompleksu ($p \in \mathbb{R}$ in $0 \leq p \leq 1$). Vrednost $F_p[i,j] = 1$, če $i$-ti simpleks vsebuje vozlišče $j$, sicer je 0. Nato izračunamo produkt redkih matrik csr($F_p$) $\times$ csr($E$). Iz vrednosti produkta določamo novonastale simplekse dimenzije $p+1$ in gradimo $(p+1)$-skeleton. Postopek rekurzivno ponavljamo.

Za naš primer je 1-skelet:

$$F_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \Rightarrow \text{csr}(F_1) = \begin{bmatrix} 0 & 0 & 1 & 1 & 2 & 2 & 3 & 3 \\ 0 & 1 & 1 & 2 & 1 & 3 & 2 & 3 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

Zatem izračunamo matrični produkt csr($F_p$) $\times$ csr($E$) ter iz njega naredimo razširjeno matriko. Vrednost v tretji vrstici matrike csr($F_1E$) nam pove dimenzijo simpleksa, ki ga gradijo simpleksi $i$ in $j$. V našem primeru

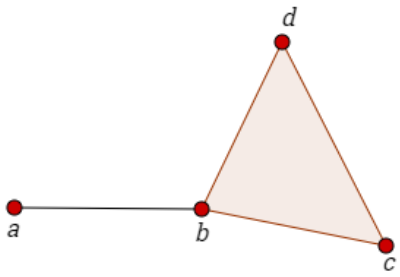
se vozliče 4 in povezava 23 povezujejo v trikotnik $\triangle 234$ (2-dimenzionalni simpleks).

$$\mathrm{csr}(F_1E) = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 2 & 2 & 3 \\ 3 & 2 & 1 & 3 & 2 & 3 & 2 & 3 \\ 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 \end{bmatrix} \Rightarrow F_1E = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Sedaj lahko izračunamo $F_2$:

$$\mathrm{csr}(F_2) = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix} \Rightarrow \mathrm{csr}(F_2) = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} \Rightarrow \mathrm{csr}(F_2E) = \begin{bmatrix} 0 & 0 \\ 3 & 2 \\ 1 & 2 \end{bmatrix}.$$

Ker je v $\mathrm{csr}(F_2E)$ matriki zopet največje število 3, pomeni, da ni več simpleksov višje dimenzije in na tej točki lahko proces gradnje simpleksov ustavimo.




Slika 3.3: Simplicialni kompleks, ki ga gradijo štirje simpleksi dimenzije 0 (vozlišča), štirje simpleksi dimenzije 1 (povezave) in en simpleks dimenzije 2 (trikotnik).

Algoritem za gradnjo VR simplicialnega kompleksa je napisan v programskem jeziku `Python` z uporabo `numpy` in `scipy` knjižnic za delo z matrikami (algoritem 1).

**Algorithm 1** Algoritem Vietoris-Rips

Vhod: $M \leftarrow$ Matrika razdalj
Vhod: $\varepsilon \leftarrow$ Prag razdalje
Vhod: $k \leftarrow$ Dimenzija, do katere gradimo simplicialne komplekse
Izhod: $VR \leftarrow$ Vietoris-Ripsov simplicialni kompleks
**procedure** VietorisRips($M$, $\varepsilon$, $k$)
    $v \leftarrow$ inicializiraj prazen niz za vozlišča
    $e \leftarrow$ inicializiraj prazen niz za povezave
    $VR \leftarrow$ inicializiraj prazen niz za simplicialne komplekse
    **if** $M[i][j] \leq \epsilon$ **then**
        $v \leftarrow$ dodaj vsa vozlišča $i$ in $j$ iz $M$
        $e \leftarrow$ dodaj vse povezave $(i, j)$ iz $M$
        $VR \leftarrow$ dodaj vozlišča $v$ v $VR$ simplicialni kompleks
    **if** $k == 0$ **then return** $VR$
    $VR \leftarrow$ dodaj povezave $e$ v $VR$ simplicialni kompleks
    **if** $k == 1$ **then return** $VR$
    **if** $(i, j) \in e$ **then**
        **if** $i < j$ **then**
            $E[i][j] = 1$
    **else**
        $E[i][j] = 0$
    $kLice \leftarrow e$
    **for all** $n \in 1..k$ **do**
        $lice \leftarrow kLice[n]$
        **if** $v \in lice$ **then**
            $F_n \leftarrow$ naredi csr matriko za $n$-skeleton
            $E[lice][v] = 1$
        $FE \leftarrow F_n$ x $E$ (zmnoži csr matrike 0-skeletona in n-skeletona)
        **if** $FE[lice][v] == n + 1$ **then**
            $kLice \leftarrow lice + v$

30: **if** $len(kLice) == 0$ **then return** $VR$
**return** $VR$

## 3.5 Filtracija in računanje vztrajne homologije

Glavni cilj vztrajne homologije je proučevanje spreminjanja topoloških značilk pri različnih resolucijah. To dosežemo tako, da simplicialni kompleks zgradimo postopno in predstavimo kot zaporedje naraščajočih simplicialnih kompleksov. Takšnemu zaporedju pravimo **filtracija**.

Filtracijom merimo življenjske cikle določene topološke značilke. Nastajanje nove topološke značilke (komponente, cikla, luknje ...) označimo kot **rojstvo**. Izginjanje topološke značilke (spajanje komponente z drugo komponento ali pa zapiranje luknje) imenujemo **smrt**. Topološke značilke, ki preživijo dlje časa, so invariante osnovnega sistema in jih obravnavamo kot pomembne. Ostale, ki živijo manj časa, obravnavamo kot šum.

V topološki analizi podatkov obstajajo utemeljitve in opisi pomena topoloških značilk do dimenzije 3. Za višje dimenzije je pomen manj jasen. V magistrski nalogi smo največ pozornosti posvetili analizi povezanih komponent (torej topoloških značilk dimenzije 0).

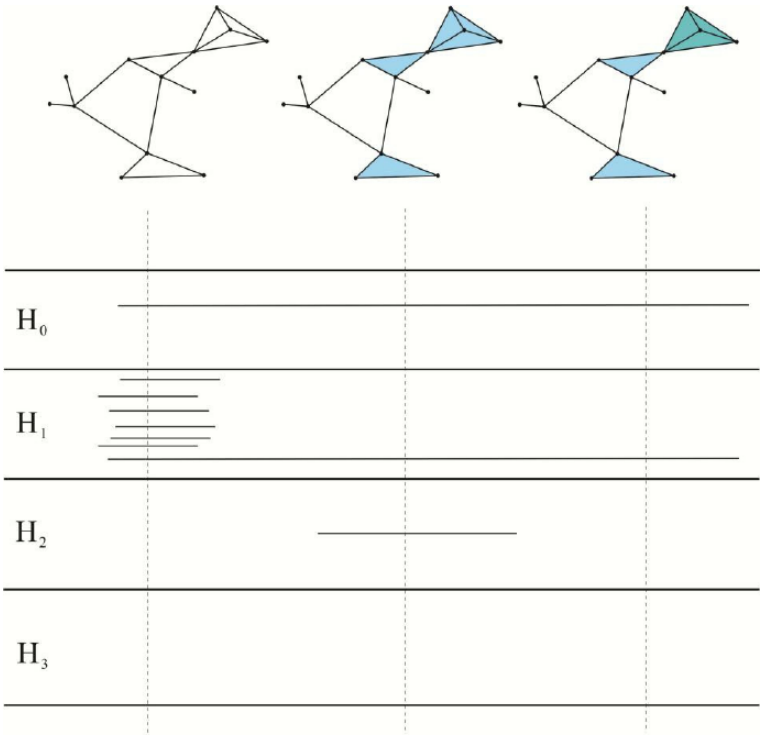
Pri izgradnji VR kompleksa metodom filtracije je zelo pomembno določiti optimalno število korakov pri filtraciji ter ustrezno razdaljo. Po eni strani veliko število korakov povzroča zelo počasno spreminjanje kompleksa. Po drugi pa se lahko pri majhnem številu korakov v enem koraku naredijo velike spremembe, ki jih ne zasledimo. Število korakov je nujno prilagoditi sami strukturi podatkov.

Kot je opisano v poglavju 3.4, smo pri gradnji filtracije z uporabo VR algoritma točke v našem primeru v topološki prostor dodajali postopoma, s spreminjanjem razdalje $\varepsilon$. Za zgornjo mejo, do katere delamo filtracijo, smo izbrali $\varepsilon$ (označimo ga z $\varepsilon_{max}$), pri katerem se je zadnja točka dodala v simplicialni kompleks. Kot optimalno število korakov za naš podatkovni nabor se je izkazalo 100 korakov filtracije. Pri manjšem številu korakov istočasno se je zgeneriralo veliko število simpleksov. Večje število korakov pa

ni vidno izboljšalo rezultatov, bilo pa je tudi precej počasnejše.

Filtracijo smo dobili tako, da smo razpon od 0 do $\varepsilon_{max}$ razdelili na 100 intervalov. Izbrali smo prvo vrednost $\varepsilon$, ki določa interval, in do te razdalje gradili VR simplicialni kompleks po algoritmu, opisanem v poglavju 3.4. V nastalem VR simplicialnem kompleksu smo prešteli topološke značilke različnih dimenzij, ki so rojene v prvem koraku filtracije.

V drugem koraku smo izbrali naslednjo vrednost $\varepsilon$ in v obstoječi simplicialni kompleks dodali nove točke in nove povezave. S ponovnim štetjem topoloških značilk smo ugotovili spremembe topološkega prostora. Pri dodajanju novih povezav se je zgodilo, da so nekatere od topoloških značilk iz predhodnega koraka izginile (umrle), medtem ko so bile rojene nekatere nove.




Slika 3.4: Postopek filtracije. Gradnja simplicialnega kompleksa čez korake filtracije (zgoraj) ter ustrezne črtne kode, ki prikazujejo rojstvo in smrt topološke značilnosti (spodaj).

Rojstvo in smrt topoloških značilk smo beležili v t.i. vztrajnem diagramu in v obliki črtnih kod [6]. Postopek smo ponavljali stokrat, oziroma dokler nismo dosegli vrednosti $\varepsilon_{max}$. Slika 3.4 prikazuje opisani postopek, le da so zaradi jasnosti vse točke prisotne od začetka.

Podan opis je posplošitev malo bolj zapletenega računskega postopka, ki je v ozadju. Standardna metoda za računanje vztrajne homologije je redukcijski algoritem. Detajlni opis in psevdokoda algoritma je podana v učbeniku A. Zomorodiana 2 [29], tukaj pa bomo uporabljali najbolj pomemben del algoritma.

**Algorithm 2** Vztrajna Homologija

Vhod: $M \leftarrow$ Matrika razdalj
Vhod: $F \leftarrow$ Koraki filtracije
Vhod: $k \leftarrow$ Maksimalna dimenzija, do katere gradimo simplicialne komplekse
Izhod: $VH \leftarrow$ Koordinate točk vztrajnega diagrama

---

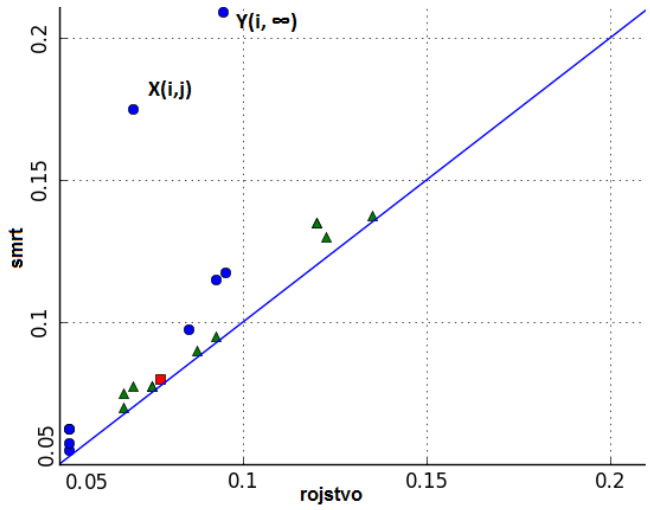
**procedure** VztrajnaHomologija($M$, $F$, $k$)
    $T \leftarrow$ Inicializiraj seznam za simplicialne komplekse
    **for all** $\varepsilon_i \in F$ **do**
        $VR \leftarrow VietorisRips(M, \varepsilon_i, k)$
        *T[i]* $\leftarrow$ Dodaj tiste simplekse iz $VR$, ki niso v $T[i-1]$
    $VH \leftarrow$ IntervalVztrajnosti(T) [29]
    **return** $VH$

## 3.6 Vztrajni diagrami in interval zaupanja

Pri računanju vztrajne homologije spreminjanje topološkega prostora vizualno prikazujemo s točkami v vztrajnem diagramu, kjer na abscisi predstavljamo rojstvo, na ordinati pa smrt topološke značilke. Točka $x(i, j)$, prikazana na sliki 3.5, predstavlja povezano komponento, ki se je rodila v nekem koraku filtracije $i$ in živela do koraka $j$. Komponente, ki preživijo do konca filtracije, označimo z $y(i, \infty)$, kjer je $i$ korak, v katerem se je komponenta

rodila. Te prikažemo na zgornjem robu vztrajnega diagrama.

Lahko rojstvo in smrt prikažemo s črtno kodo, kjer dolžina črte predstavlja preživetje topološke značilke skozi korake filtracije (slika 3.4 – rdeče in zelene črte). Trenutek rojstva in smrti razberemo s časovne premice. Ker so vztrajni diagrami v našem primeru bili bolj primerni za analizo, se bomo v nadaljnjem opisu večinoma osredotočili nanje.




Slika 3.5: Vztrajni diagram. Na $x$-osi, ki predstavlja rojstvo, in $y$-osi, ki predstavlja smrt topološke značilnosti, smo prikazali intervale filtracije (razdalje). Modre pike predstavljajo topološke značilnosti dimenzije 0, zeleni trikotniki pa predstavljajo topološke značilnosti dimenzije 1. Rdeči kvadrat je topološka značilnost dimenzije 2.

Do sedaj smo večkrat omenili, da topološke značilke kratkega življenja obravnavamo kot šum in jih zanemarimo. V vztrajnem diagramu so te značilke natanko tiste točke, ki so zelo blizu diagonali vztrajnega diagrama (v črtni kodi so to zelo kratke črtice). Vprašanje, ki se je zastavilo, je, kako določiti, kaj je ,,dovolj blizu" ter določiti skupine bolnikov, ki so bolj ločene druga od druge, in hkrati poiskati manjše podskupine v določenih skupinah. V tem primeru je bilo nujno pravilno ločiti šumne od pomembnih značilk.

V ta namen smo implementirali algoritem, ki računa interval zaupanja na vztrajnih diagramih. Algoritem temelji na izreku o stabilnosti vztrajnega diagrama. Za vztrajni diagram pravimo, da je **stabilen**, če majhne spremembe v podatkih povzročajo majhne spremembe v vztrajnem diagramu [4]. Psevdokoda opisanega algoritma je podana kot algoritem 3.

Stabilnost določimo z uporabo t.i. razdalje ozkega grla (ang. *bottleneck distance*) med diagrami, ki je definirana s predpisom:

$$W_\infty(X_p, Y_p) = |x - y|_\infty. \tag{3.4}$$

$$|x - y|_\infty = max\{|x_1 - y_1|, |x_2 - y_2|\}, \tag{3.5}$$

kjer so $x$ in $y$ točke vztrajnih diagramov $X_p$ oziroma $Y_p$.

Za ocenjevanje izračunanih homoloških značilk množice $S$ ter določanja intervala zaupanja namesto celotne množice podatkov $S$ raje izberemo dovolj velik vzorec $M$, ki podobno opisuje topološki prostor podatkov kot množica $S$. Interval zaupanja $C$ je potem podmnožica vseh vztrajnih diagramov $X_{M_i}$, pri katerih je razdalja od diagrama $X_S$ največ $c$, kjer je $c$ prag intervala zaupanja [7]:

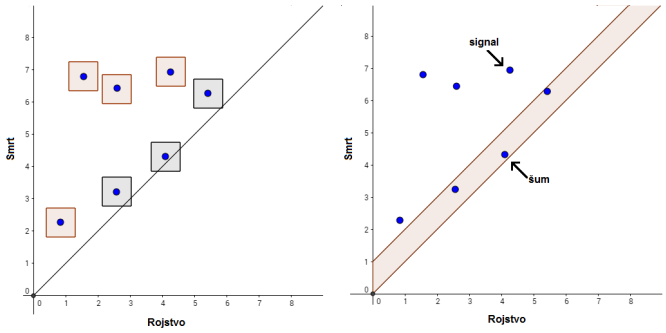
$$C = \{X_{M_i} : W_\infty(X_S, X_{M_i}) \leq c\}. \tag{3.6}$$

Grafično lahko to predstavimo na vztrajnem diagramu z risanjem centriranih kvadratkov stranice $2c$ okoli vsake točke na diagramu. Točko $p$ obravnavamo kot topološki šum, če njen ustrezni kvadrat seka diagonalo. Na drug način lahko interval zaupanja predstavimo tudi s trakom širine $\sqrt{2c}$ nad diagonalo (slika 3.6).

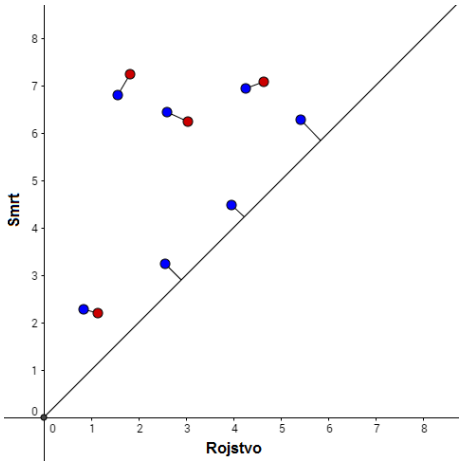
Pri določanju intervala zaupanja oziroma za računanje razdalje med diagrami moramo najprej povezati ustrezne točke vztrajnih diagramov. Točke, ki nimajo para, je potrebno povezati z najbližjo točko na diagonali, kot je prikazano na sliki 3.7.

Pri reševanju problema smo se poslužili algoritmov, ki spadajo v področje diskretne matematike, in sicer algoritmov za maksimalno prirejanje (ang.

*matching*). Točke vztrajnih diagramov smo preslikali v dvodelni graf, nato pa smo poiskali maksimalno prirejanje z uporabo algoritma za maksimalni pretok (Ford Fulkersonov algoritem).




Slika 3.6: Interval zaupanja. Na sliki levo so opisani intervali zaupanja prikazani s kvadratom stranice $2c$, narisanim okoli točke v vztrajnem diagramu. Če kvadrat seka diagonalo, potem točko, ki pripada temu kvadratu, obravnavamo kot šum. Slika desno je drugi način prikazovanja intervala zaupanja. Nad diagonalo narišemo trak širine $\sqrt{2c}$ . Točke, ki so v traku, obravnavamo kot šum.




Slika 3.7: Prirejanje. Slika prikazuje način prirejanja točk dveh vztrajnih diagramov (modre in rdeče točke). Točke, ki nimajo para, priredimo najbližji točki na diagonali.

**Algorithm 3** Interval Zaupanja

Vhod: $M \leftarrow$ Matrika razdalj
Vhod: $F \leftarrow$ Koraki filtracije
Vhod: $k \leftarrow$ Maksimalna dimenzija, do katere gradimo simplicialne komplekse
Izhod: $X_{MC} \leftarrow$ Vztrajni diagram z intervalom zaupanja

**procedure** IntervalZaupanja($M$, $F$, $k$)
    $VH \leftarrow$ VztrajnaHomologija(M, F, k)
    $X_M \leftarrow$ Nariši vztrajni diagram iz VH
    $C_n \leftarrow$ Inicializiraj niz za interval zaupanja
    $n \leftarrow$ Določi število ponovitev
    **while** $i \leq n$ **do**
        $m \leftarrow$ Izberi naključni vzorec iz M
        $vh \leftarrow$ VztrajnaHomologija(m, F, k)
        $X_m \leftarrow$ Nariši vztrajni diagram iz $vh$
        $MP \leftarrow$ MaksimalnoPrirejanje($X_M$, $X_m$)
        $W_\infty(X_M, X_m) \leftarrow$ RazdaljaOzkegaGrla(MP)
        $C_n \leftarrow W_\infty(X_M, X_m)$
    $c \leftarrow$ Določi prag intervala zaupanja
    $C_n \leftarrow$ Sortiraj niz $C_n$
    $C \leftarrow$ Izberi podmnožico $C_i$ iz $C_n$ tako da je $C_n[i] \leq$ c
    $X_{MC} \leftarrow$ Prikaži interval zaupanja C na vztrajnem diagramu $X_M$
    **return** $X_{MC}$

# Poglavje 4

# Rezultati in razprava

V tem poglavju bomo prikazali rezultate, pridobljene z uporabo topoloških algoritmov. Opisu rezultatov bo sledila ustrezna grafična predstavitev.

## 4.1 Potek analize in testiranje metode

Po uspešni implementaciji naštetih algoritmov smo pristopili k analizi podatkov. Potek analize je sledil korakom, prikazanim v algoritmu 4.

Pred začetkom analize podatkov, opisanih v poglavju 3, smo naredili nekaj testnih preizkusov zaradi preverjanja in evalvacije metode. Delovanje algoritma smo najprej testirali na množici, ki vsebuje podatke o genski izraženosti zdravega ter podatke o genski izraženosti rakavega tkiva. V eni od novejših raziskav so I. Kosti in sodelavci [11] pokazali, da se genska izraženost genov v zdravem tkivu značilno razlikuje od izraženosti genov v bolnem tkivu. Znanstveniki so dokazali, da so geni zdravega tkiva bolj korelirani od genov bolnega tkiva. Glede na to smo pričakovali, da bo implementirani algoritem ločil podatke na dve skupini.

Podatkovni nabor je vseboval 130 vzorcev, od tega 67 iz rakavega in 63 iz zdravega tkiva. V analizo je bilo uvrščenih 15.658 genov. Pri prvem preizkusu je algoritem uspešno razdelil podatke na dve skupini, kar prikazuje slika 4.1. Želeli smo potrditi, da rezultat algoritma ni naključje, zato smo testiranje

algoritma ponovili večkrat, tudi na različnih množicah. Rezultat testiranja je bil pozitiven.

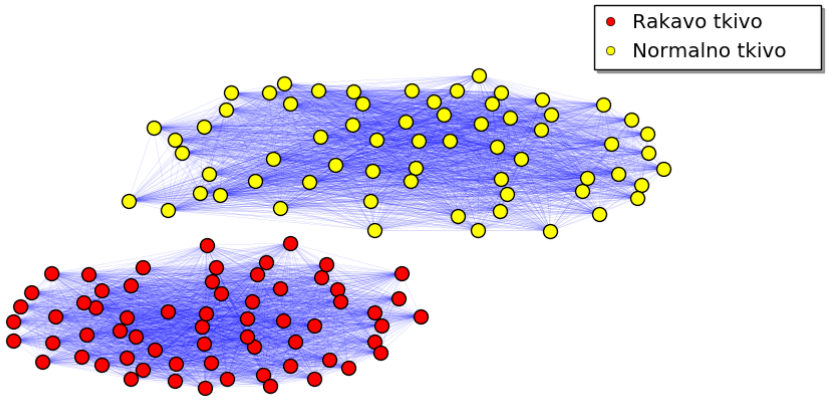
**Algorithm 4** Topološka analiza podatkov

Vhod: $ICGC \leftarrow$ Podatki iz podatkovne zbirke ICGC
Izhod: $G \leftarrow$ Gruče donatorjev
Izhod: $KM \leftarrow$ Krivulje preživetja
Izhod: $ss \leftarrow$ Silhouette ocena
Izhod: $X_{MC} \leftarrow$ Vztrajni diagram z intervalom zaupanja

**procedure** TDA($ICGC$)
    $D \leftarrow$ Preberi podatke iz $ICGC$
    $I \leftarrow$ Shrani metapodatke
    $M \leftarrow$ Izračunaj matriko razdalj
    $\varepsilon \leftarrow$ Določi vrednost za maksimalno razdaljo
    $F \leftarrow$ Določi korake filtracije od 0 do $\varepsilon$
    $k \leftarrow$ Določi maksimalno dimenzijo za gradnjo simplicialnih kompleksov
    $X_{MC} \leftarrow$ IntervalZaupanja($M$, $F$, $k$)
    $G \leftarrow$ Naredi gruče od topoloških značilnosti dimenzije 0
    **for** $g \in G$ **do**
        $ss \leftarrow$ SilhouetteScore($g$)
        $KM \leftarrow$ Preživetje($g$)
    $G \leftarrow$ Priredi metapodatke iz $I$
    **return** $G$, $KM$, $ss$, $X_{MC}$

Uspešnost metode je posledica tega, da nam za razliko od ostalih metod, ki iščejo vzorec v podatkih, topološka analiza daje ,,opis" podatkov. Kot prost primer lahko vzamemo množico točk, ki predstavljajo elipso. Če na primeru, kadar je ena os elipse veliko daljša kot druga (sploščena elipsa), uporabimo metodo analize glavnih komponent ali večdimenzionalno skaliranje, bomo dobili množico točk, ki predstavljajo premico (eno skupino). Z uporabo topološke metode bomo na istem primeru prav tako dobili eno

povezano komponento (skupino) in en cikel. Poglejmo še drugi primer, in sicer kadar imamo ,,normalno" elipso. V tem primeru bi z metodo glavnih komponent dobili dve množici točk, ki opisujeta dve premici (dve skupini). Topološka metoda bi tudi v tem primeru dala eno povezano komponento in en cikel.




Slika 4.1: Delitev množice na dve skupini, z vzorci zdravega tkiva in vzorci rakavega tkiva.

## 4.2 Rezultati analize

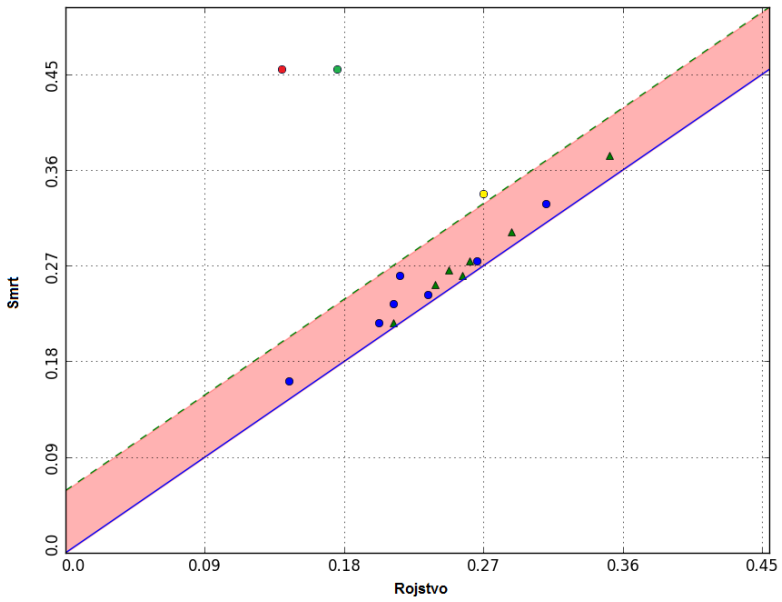
Po uspešnem testiranju smo pristopili k analizi predhodno opisanih podatkov, in sicer raka prsi – BRCA, raka jajčnikov – OV in raka pljuč – LUSC. Rezultati so podani v nadaljevanju.

### 4.2.1 Rak prsi

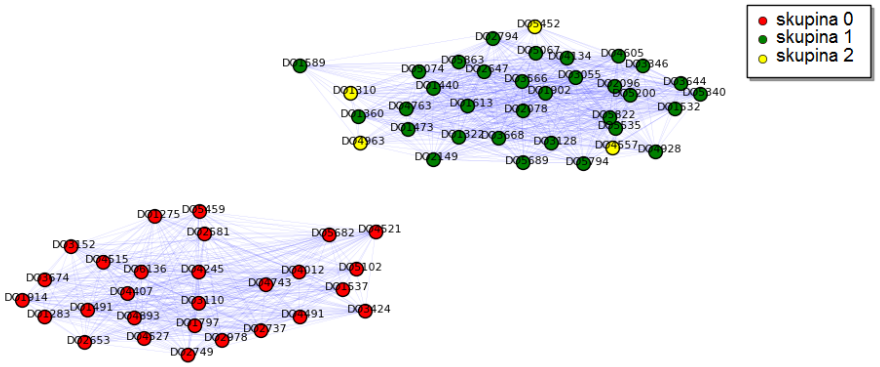
Podatkovna množica prsnega raka je vsebovala 65 bolnikov. V analizo smo uvrstili 17.662 genov, med katerimi je bilo veliko takih, za katere je že znano, da prispevajo k zaviranju napredovanja raka prsi. Med njimi so bili denimo znani tumorski zaviralci raka prsi, BRCA1 in BRCA2, ter geni

PIK3CA, TP53, TTN, TTN-AS1, RP11, PCDHGA1, PCDHGA2, PCDHGA3, PCDHA1, CDH1, ki so po ICGC analizi najpogostejše mutirani geni v tem tipu raka.

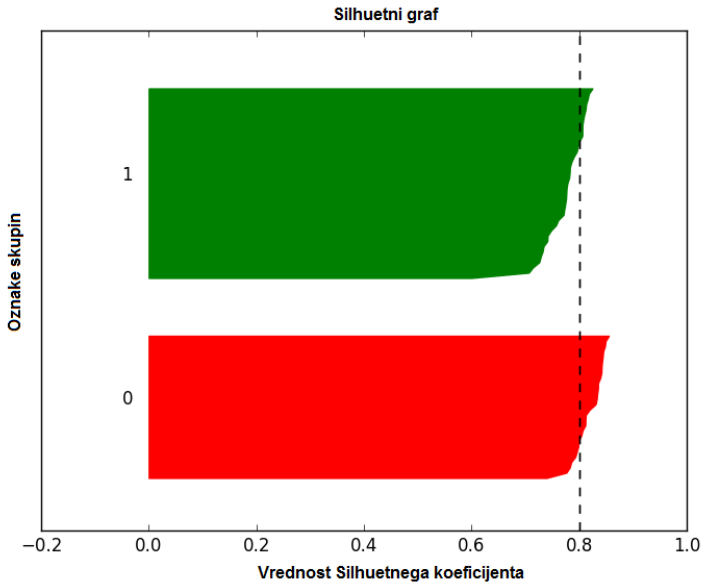
Z analiziranjem podatkovne množice z uporabo topološkega algoritma ter glede na vztrajni diagram (slika 4.2) smo ugotovili, da obstajata dva podtipa raka prsi (slika 4.3). Kot je razvidno iz vztrajnega diagrama, rdeča in zelena točka predstavljata dve topološki značilnosti dimenzije 0 oziroma povezani komponenti, ki ,,vztrajata" do konca filtracije. Prikazani točki barvno ustrezata dvema skupinama raka prsi.




Slika 4.2: Vztrajni diagram, pridobljen z analizo raka prsi. Na $x$-osi je prikazano rojstvo, na $y$-osi pa smrt topološke značilnosti. Vrednosti na oseh $x$ in $y$ predstavljajo razdalje, pri katerih gradimo simplicialne komplekse tekom filtracije. Na primer, povezana komponenta, obarvana rumeno, se je začela pri razdalji 0.27 in končala pri razdalji 0.34. Rdeči trak predstavlja interval zaupanja.




Slika 4.3: Razvrstitev raka prsi z uporabo vztrajne homologije. Na sliki sta prikazani dve ločeni skupini raka prsi (rdeča in zelena) ter podskupina, ki predstavlja t.i. osamelce (rumena).



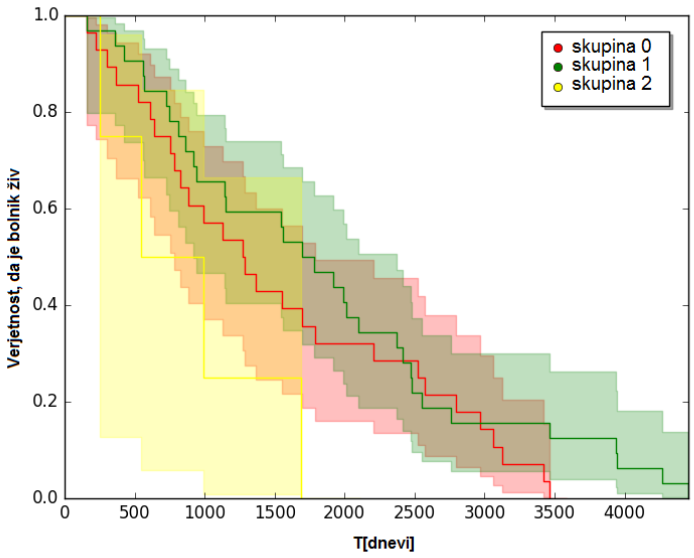

Slika 4.4: Silhuetni graf za razvrstitev vzorcev raka prsi. Barvi Silhuetnega grafa ustrezata barvam pridobljenih skupin, debelina pa velikosti skupine. Prekinjena črta predstavlja povprečni Silhuetni koeficient.

Veliko predhodnih raziskav je tudi potrdilo, da obstajata dva tipa raka

prsi, in sicer bazalni in luminalni. Za bazalni je karakteristična mutacija gena BRCA1, medtem ko imamo pri luminalnem bolj prisotno mutacijo gena BRCA2.

Rezultat razvrščanja smo ocenili z Silhuetnim koeficientom, ki v tem primeru znaša 0,81 (slika 4.4), na lestvici od -1 do 1. Glede na to, da večja vrednost pomeni boljšo delitev, lahko rečemo, da je implementirana metoda dala zelo dobre rezultate v primeru raka prsi.

Z analizo preživetja bolnikov smo z uporabo Kaplan-Meierjeve metode (slika 4.5) potrdili, da ima skupina bolnikov daljši čas preživetja (zelena krivulja), in sicer več kot 4.500 dni od rdeče skupine (rdeča krivulja), kjer je čas preživetja manj kot 3.500 dni.




Slika 4.5: Krivulja preživetja za vsako skupino raka prsi posebej. Na $x$-osi je prikazan vremenski razpon. Na $y$-osi je prikazana verjetnost, da bolnik še zmeraj živi. Barve krivulj ustrezajo barvam pridobljenih skupinah.

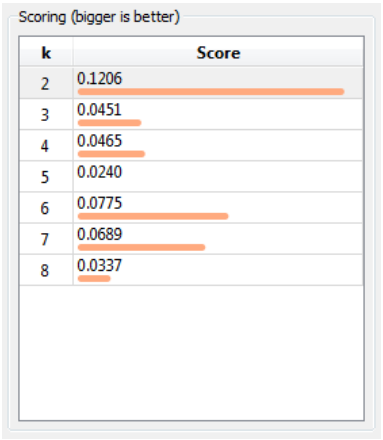
Z računanjem intervala zaupanja za vztrajni diagram (slika 4.2 – rdeči trak) smo pokazali, da ena povezana komponenta ,,vztraja" samostojno več

časa (slika 4.2 – rumena točka). To komponento sestavljajo samo štirje bolniki. Za njih ne moremo reči, da predstavljajo podskupino zase. To smo potrdili tudi z računanjem Silhuetnega koeficienta, ki znaša 0,485. Z analiziranjem preživetja teh bolnikov smo odkrili, da je njihov čas preživetja veliko krajši (slika 4.5 – rumena krivulja), torej nekaj več kot 1.500 dni, kar je veliko manj kot čas preživetja ostalih bolnikov v skupini, kateri pripadajo (slika 4.3 – zelena skupina). Za njih lahko rečemo, da predstavljajo t.i. osamelce.

Že pri prvi analizi smo pokazali, da je vztrajna homologija precej uspešna v analizi bioloških podatkov. Videli smo, da deluje na zelo velikih dimenzijah, ter da uspešno obravnava osamelce.

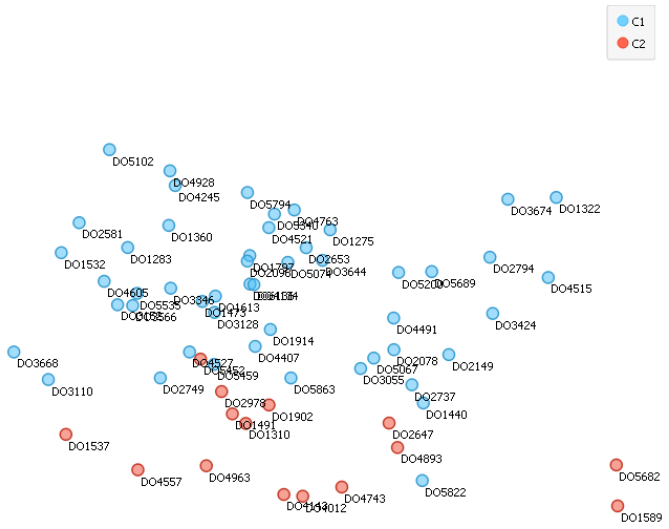
Isto podatkovno množico smo analizirali z metodo razvrščanja z voditelji ter metodo hierarhičnega razvrščanja, ki jih ponuja paket `Orange`[1].

Metoda z voditelji nam je kot rezultat z najboljšim Silhuetnim koeficientom ponujala dve gruči (slika 4.6). V primerjavi z našim algoritmom je Silhuetni koeficient zelo nizek, znaša samo 0,1206 (slika 4.7). Glede na to, da metoda z voditelji uporablja evklidsko razdaljo, ki ni najbolj primerna za našo podatkovno množico, boljši rezultat ni bil pričakovan.



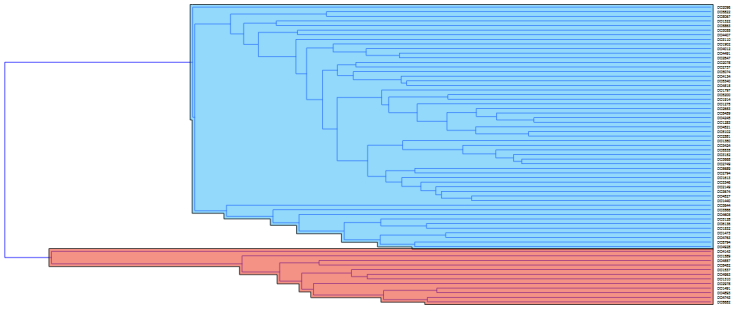

Slika 4.7: Silhuetni koeficienti pri iskanju optimalnega števila skupin za metodo z voditelji.

[1] `http://orange.biolab.si`

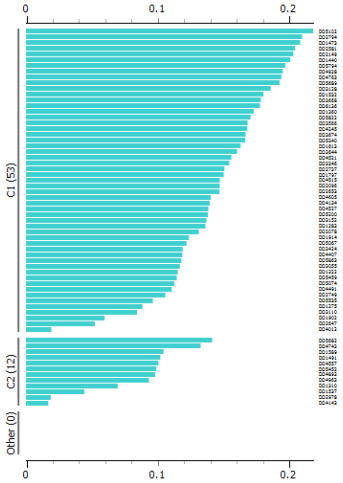



Slika 4.6: Podskupine raka prsi, pridobljene z uporabo metode voditeljev.

Metoda hierarhičnega razvrščanja je dala nekaj boljših rezultatov. To je bilo pričakovano, saj smo v primeru hierarhičnega razvrščanja za razdaljo uporabljali Pearsonov koeficient. Kot rezultat hierarhičnega razvrščanja z uporabo utežene metode smo dobili dve podskupini (slika 4.8). Tudi Silhuetni koeficient je v primeru hierarhičnega razvrščanja boljši (slika 4.9).



Slika 4.8: Podskupine raka prsi, pridobljene metodom hierarhičnega razvrščanja.




Slika 4.9: Silhuetni graf razvrščanja raka prsi, pridobljen metodom hierarhičnega razvrščanja.

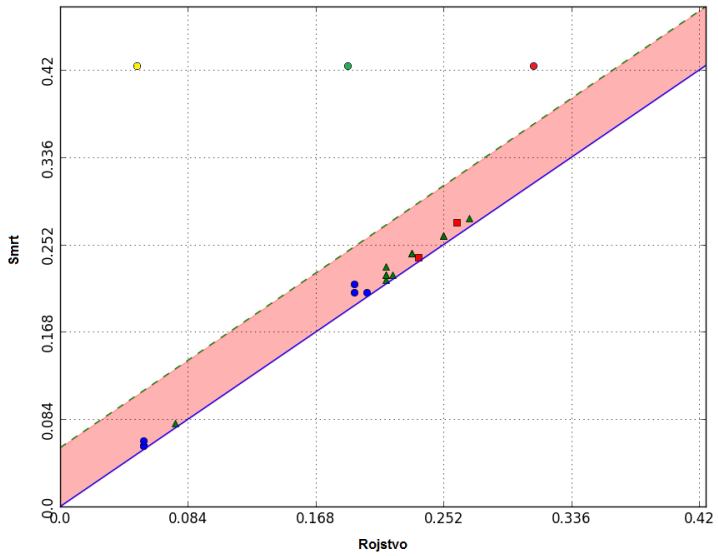
### 4.2.2 Rak jajčnikov

Rak jajčnikov velja za sedmo najpogostejšo obliko raka pri ženskah v svetu.[2] Rak jajčnikov lahko nastane iz vseh treh vrst celic, ki sestavljajo jajčnik: iz

[2] `http://www.cancerresearchuk.org`

epitelijskih, kličnih celic in celic strome. Najpogostejši so epitelijski tumorji, ki predstavljajo več kot 90% vseh primerov. Redkejša je skupina neepitelnih tumorjev, kamor poleg stromalnih in kličnih (germinativnih) tumorjev prištevamo tudi mešane maligne Muelerjeve tumorje jajčnikov.

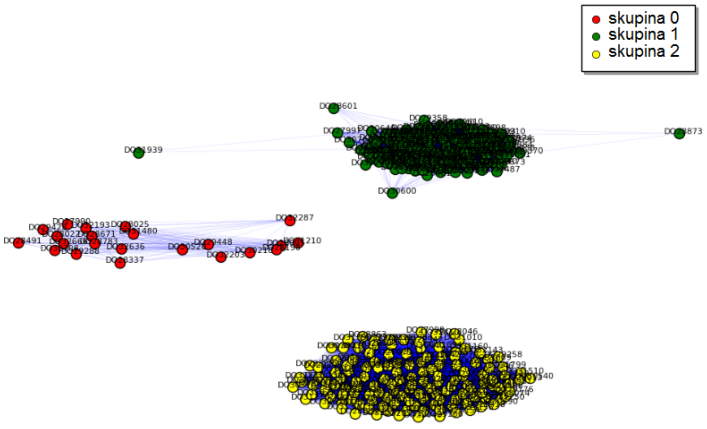
V naši raziskavi smo obravnavali skupino, ki vsebuje 292 bolnikov. V analizo je bilo uvrščenih 11.876 genov. Med njimi je TP53, ki je po ICGC najbolj pogosto mutiran gen tega tipa raka. V raziskavo smo uvrstili tudi gene BRCA1 in BRCA2, ki so poleg raka prsi zelo pogosto prisotni v raku jajčnikov.



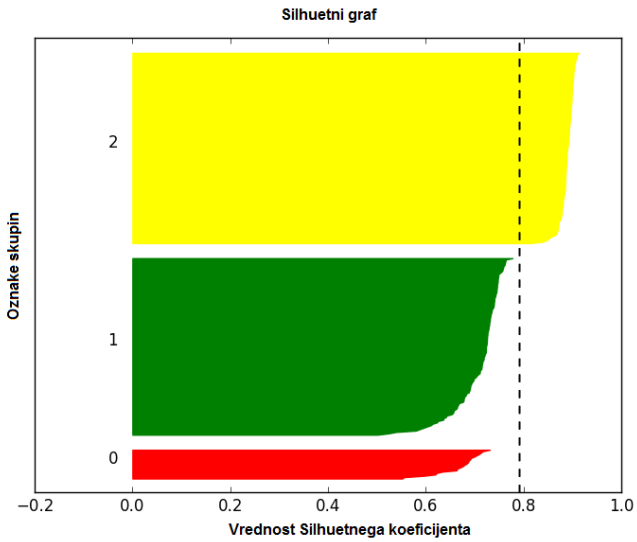

Slika 4.10: Vztrajni diagram, pridobljen s topološko analizo raka jačnikov. Na $x$-osi je prikazano rojstvo, na $y$-osi pa smrt topološke značilnosti. Vrednosti na $x$ in $y$ oseh predstavljajo razdalje, pri katerih gradimo simplicialne komplekse tekom filtracije. Rdeči trak predstavlja interval zaupanja.

Z uporabo topoloških algoritmov smo opisano množico bolnikov razvrstili v tri skupine (slika 4.11). Omenjene skupine določajo rumena, zelena in rdeča točka vztrajnega diagrama (slika 4.10).

Silhuetni koeficient v primeru razvrščanja raka jajčnikov znaša 0,79 (slika 4.12).

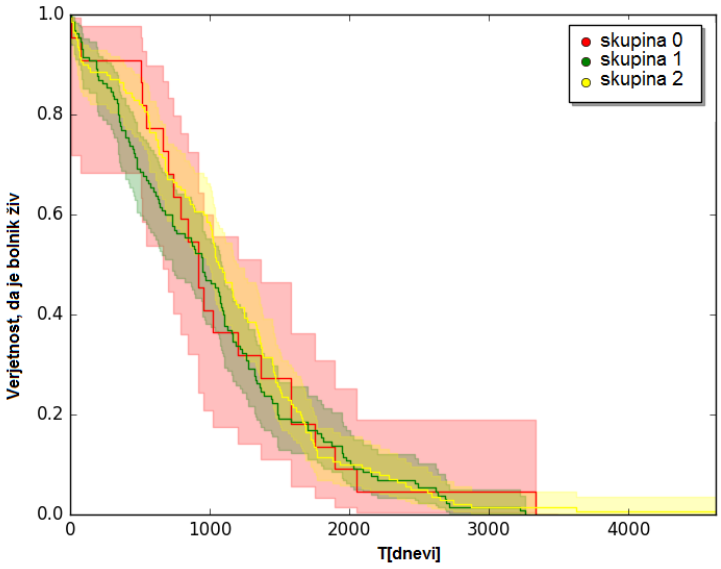



Slika 4.11: Podskupine raka jajčnikov, pridobljene s topološko analizo.




Slika 4.12: Silhuetni graf podskupin raka jačnikov. Barve Silhuetnega grafa ustrezajo barvam pridobljenih skupin, debelina pa velikosti skupine. Prekinjena črta predstavlja povprečni Silhuetni koeficient.

V primeru raka jajčnikov imata dve skupini (zelena in rdeča) približno enak čas preživetja, in sicer okoli 3.200 dni. Preživetje rumene skupine je večje od 4.000 dni (slika 4.13).




Slika 4.13: Krivulje preživetja bolnikov z rakom jačnikov. Na $x$-osi je prikazan vremenski razpon. Na $y$-osi je prikazana verjetnost, da bolnik še zmeraj živi. Barve krivulj ustrezajo barvam pridobljenih skupin raka jačnikov.

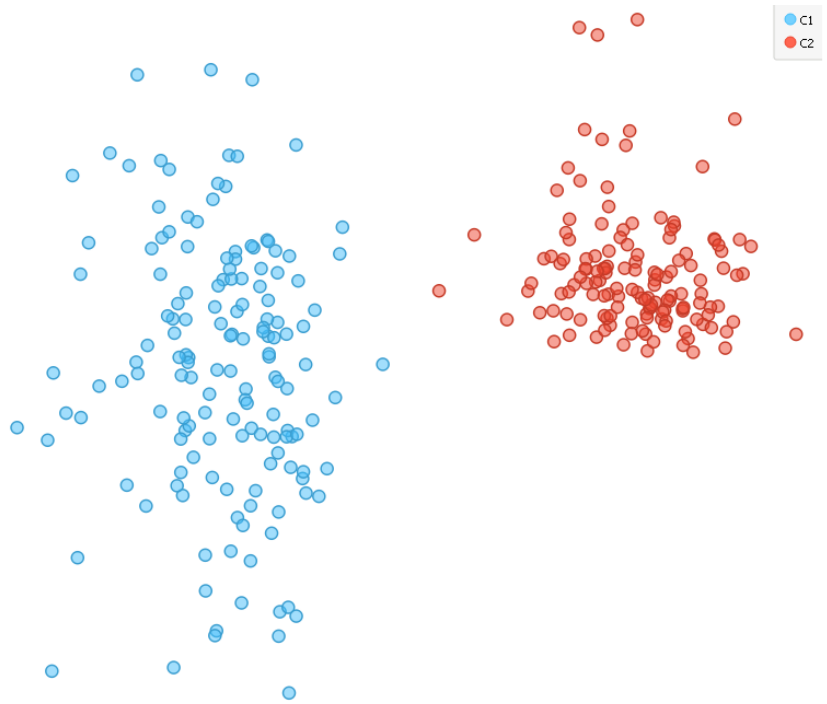
V primerjavi z našo metodo je metoda z voditelji z najboljšo vrednostjo Silhuetnega koeficienta tj. – 0,83 (slika 4.14), bolnike razvrstila v dve skupini. Skupine so jasno ločene druga od druge, kar ni bilo pričakovano, zopet zaradi uporabe evklidske razdalje.

Metoda hierarhičnega razvrščanja je – podobno kot metoda z voditelji – množico ločila na dve skupini (slika 4.16). Silhuetni koeficient je prikazan na sliki 4.17.

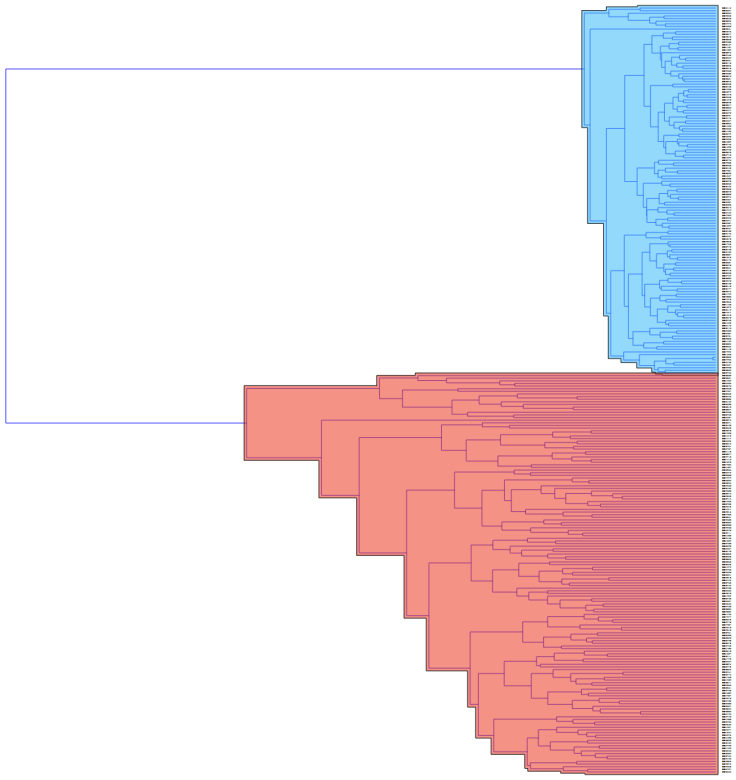
Scoring (bigger is better)

| k | Score |
|---|---|
| 2 | 0.83 |
| 3 | 0.44 |
| 4 | 0.06 |
| 5 | 0.05 |
| 6 | 0.05 |
| 7 | 0.05 |
| 8 | 0.03 |

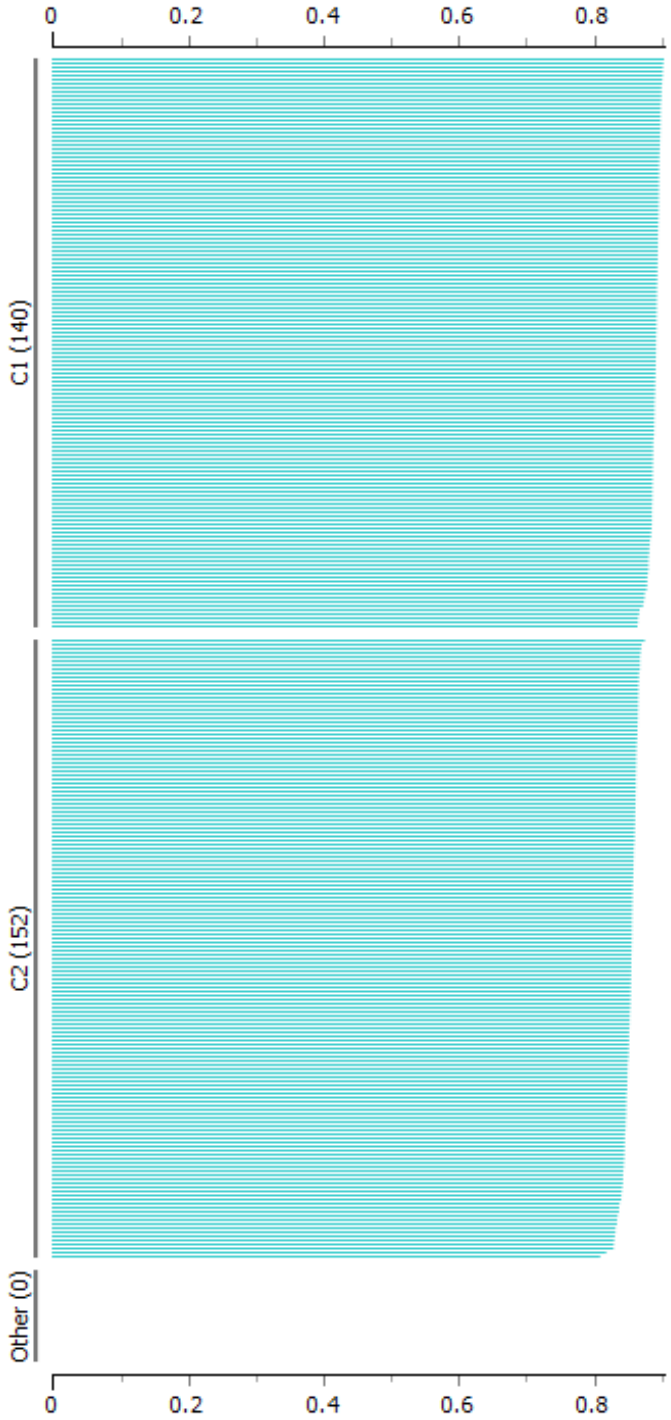
Slika 4.14: Silhuetni koeficienti pri iskanju optimalnega števila skupin za metodo voditeljev.




Slika 4.15: Razvrščanje raka jajčnikov z uporabo metode voditeljev.



Slika 4.16: Razvrščanje raka jajčnikov z uporabo hierarhičnega razvrščanja.

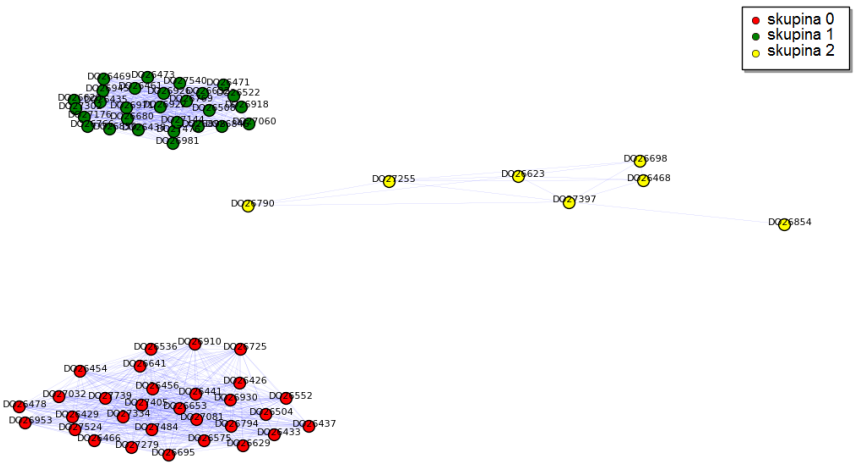



Slika 4.17: Silhuetni graf pri hierarhičnem razvrščanju.

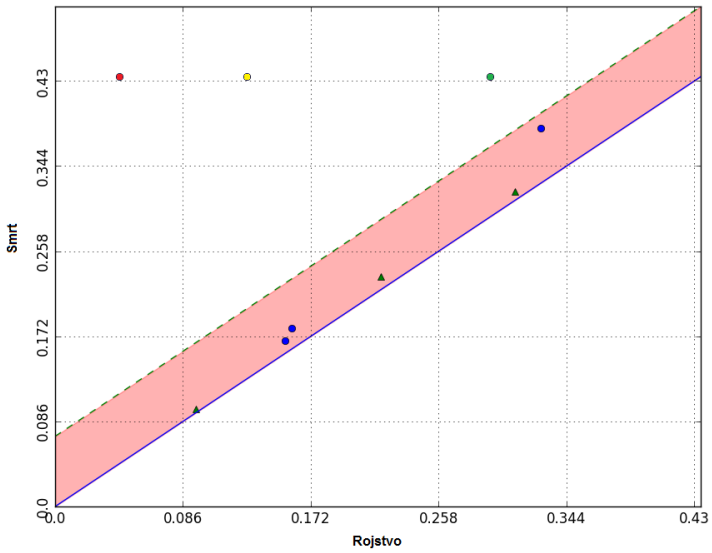
### 4.2.3 Pljučni rak

Pljučni rak spada med najpogostejša rakava obolenja. Za to vrsto raka je značilna hitra rast malignih celic. Tumorske celice se hitro širijo v okolico ter s krvjo prenašajo na ostale dele telesa. V magistrski nalogi smo – namesto analize vseh znanih tipov raka pljuč – analizirali množico od 65 bolnikov s ploščatoceličnim rakom pljuč. Pri analizi smo upoštevali 11.862 genov. Wilkerson in soavtorji [25] so v svoji raziskavi pokazali, da obstajajo trije podtipi ploščatoceličnega raka pljuč, in sicer bazalni, sekretorni in klasični podtip.

V naši analizi smo na osnovi podatkov dobili tudi tri podskupine ploščatoceličnega pljučnega raka (slika 4.18). Zaradi pomanjkljivosti kliničnih podatkov o bolnikih, ni bilo možno dokazati ali skupine uztrezajo podtipom. Vsaka od točk vztrajnega diagrama (slika 4.19), rumena, zelena in rdeča, ustrezajo skupinam na sliki 4.18.



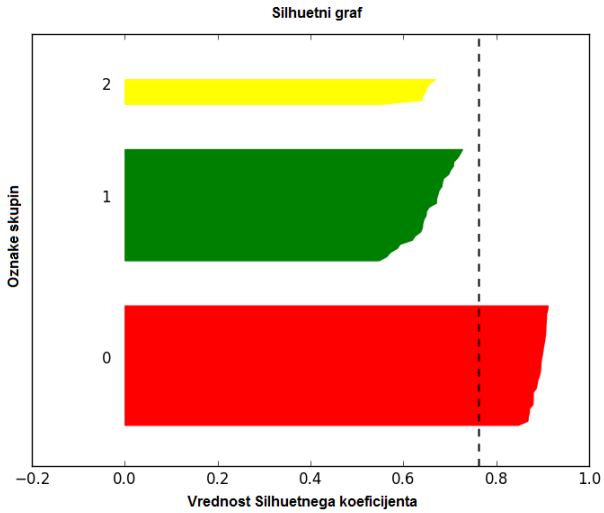

Slika 4.18: Podskupine ploščatoceličnega raka pljuč, pridobljene s topološko analizo.



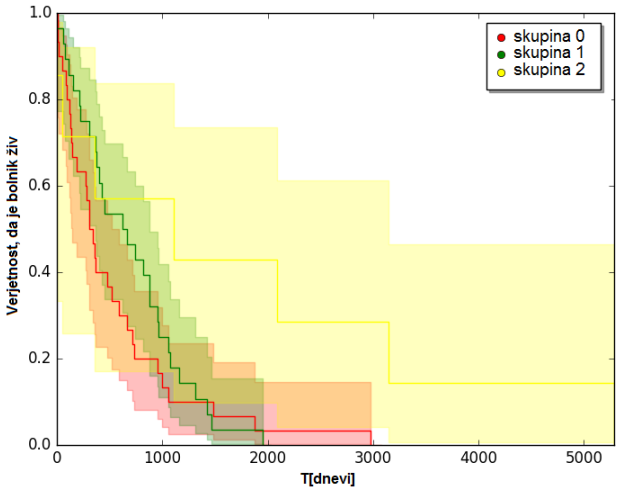

Slika 4.19: Vztrajni diagram, pridobljen s topološko analizo ploščatoceličnega raka pljuč. Na $x$-osi je prikazano rojstvo, na $y$-osi pa smrt topološke značilnosti. Vrednosti na $x$ in $y$ oseh predstavljajo razdalje, pri katerih gradimo simplicialne komplekse tekom filtracije. Rdeči trak predstavlja interval zaupanja. Tri topološke značilnosti, ki preživijo do konca filtracije, so prikazane z rdečo, rumeno in zeleno točko na diagramu.

Razvrščanje je ocenjeno z uporabo Silhuetnega koeficienta z vrednostjo 0,78 (slika 4.20).

V primerjavi s preživetjem bolnikov z rakom prsi (slika 4.5) ter preživetjem bolnikov z rakom jajčnikov (slika 4.13), lahko pri pljučnem raku opazimo, da krivulje preživetja padajo nekajkrat hitreje (slika 4.21). Možni razlog bi bil, da ima več kot polovica bolnikov po odkritju ploščatoceličnega pljučnega raka že oddaljene metastaze. Zatorej je povprečno preživetje bolnikov z rakom tega tipa okoli 5 let [20]. Skupina (slika 4.18, skupina 2), katera preživi več časa, vsebuje manjše število bolnikov. To zopet potrjuje ugotovitev, da ima zelo majhen odstotek bolnikov, obolelih za ploščatoceličnim pljučnim rakom, možnost preživetja več let.

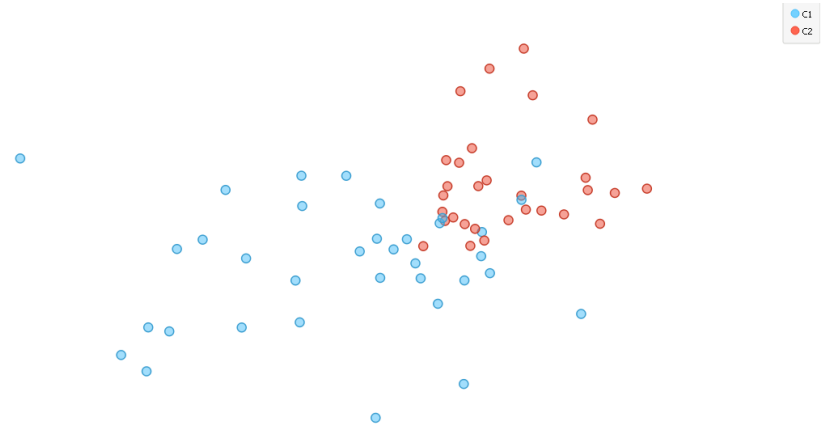



Slika 4.20: Silhuetni graf podskupin ploščatoceličnega raka pljuč. Barve Silhuetnega grafa ustrezajo barvam pridobljenih skupin, debelina pa velikosti skupine. Prekinjena črta predstavlja povprečni Silhuetni koeficient.

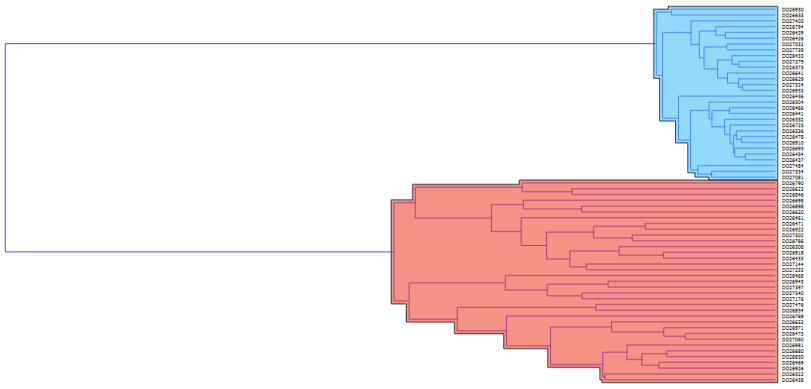



Slika 4.21: Krivulje preživetja bolnikov s ploščatoceličnim rakom pljuč. Na $x$-osi je prikazan razpon preživetja. Na $y$-osi je prikazana verjetnost, da bolnik živi. Barve krivulj ustrezajo barvam skupin raka pljuč.

V primerjavi z našo metodo sta metodi z voditelji ter metoda hierarhičnega razvrščanja oba tipa bolnikov razvrstili na dve skupini (slika 4.22 in slika 4.23).



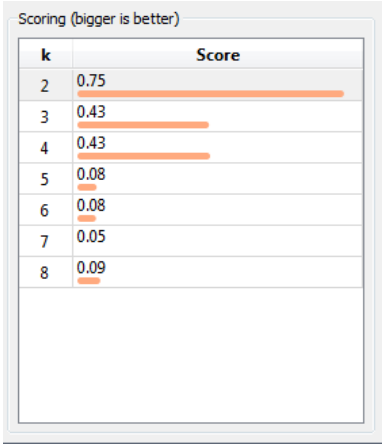

Slika 4.22: Razvrščanje ploščatoceličnega raka pljuč z uporabo metode voditeljev.



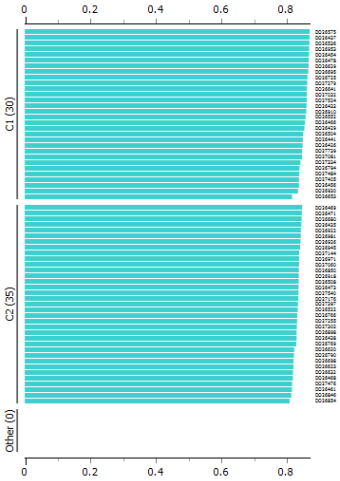
Slika 4.23: Razvrščanje ploščatoceličnega raka pljuč z uporabo hierarhične metode.

Silhuetni koeficient za metodo z voditelji je v tem primeru malenkost slabši kot pri naši metodi in znaša 0,75 (slika 4.24). Glede na Silhuetni

graf (slika 4.25) se zdi, da ima metoda hierarhičnega razvrščanja približen Silhuetni koeficient kot naša metoda.




Slika 4.24: Silhuetni koeficienti pri iskanju optimalnega števila skupin za metodo voditeljev.




Slika 4.25: Silhuetni graf pri hierarhičnem razvrščanju.

# Poglavje 5

# Sklepne ugotovitve in bodoče raziskave

Topološka analiza podatkov se je v zadnjih časih izkazala za zelo uspešna pri odkrivanju informacij v kompleksnih bioloških podatkih. Pri izdelavi magistrske naloge smo se odločili, da na podatkih o genski izraženosti rakavega tkiva uporabimo metode, ki temeljijo na računski topologiji. Cilj magistrske naloge je bil, da z uporabo vztrajne homologije na omenjenih podatkih poskusimo določiti skupine v posameznem tipu raka in napovedati preživetje bolnikov v dobljenih skupinah. Podatki, ki smo jih analizirali, so vzeti iz prosto dostopne podatkovne baze ICGC.

Glavna ideja je bila, da na podatkih o genski izraženosti rakavega tkiva postopoma gradimo Vietoris-Ripsove simplicialne komplekse ter računamo vztrajno homologijo, ki nam bo opisala ,,obliko" podatkov. Na podlagi vztrajne homologije smo izrisovali vztrajne diagrame. Z nadaljnjo analizo topoloških značilnosti dimenzije 0, prikazanih na vztrajnih diagramih, smo tako uspešno naredili razvrstitev bolnikov na skupine. Nato smo z uporabo metode Kaplan-Meier napovedali preživetje bolnikov v skupinah.

Postopek smo ponovili za tri tipe raka, in sicer za rak prsi, rak jajčnikov in rak pljuč. Tekom analize raka prsi smo ugotovili, da obstajata dva podtipa tega raka. Z ocenjevanjem dobljenih rezultatov s Silhuetnim koeficientom in

primerjanjem z drugimi metodami nenadzorovenega učenja smo pokazali, da je razvita metoda bila bolj uspešna. Rezultate je bilo potrebno tudi klinično ovrednotiti, kar žal ni bilo možno zaradi pomanjkljivosti kliničnih podatkov o bolnikih.

S topološko analizo raka jajčnikov smo bolnike razvrstili na tri skupine. Z uporabo metode voditeljev in metode hierarhičnega razvrščanja smo v primeru tega tipa raka dobili dve podskupini.

Za razliko od raka prsi in raka jajčnikov smo se v primeru pljučnega raka odločili za analizo enega od podtipov tega raka, in sicer za ploščatoceličnega raka pljuč. Na ta način smo pokazali, da lahko isto metodo uporabimo tudi v primeru odkrivanja podskupin v enem podtipu raka. Poleg tega smo dokazali, da metoda uspešno deluje na visokodimenzionalnih podatkih. V primerjavi z ostalimi metodami razvrščanja, ki poskusijo določiti vzorec v podatkih, nam topološka metoda da ,,opis", iz katerega lahko odkrijemo pomembne informacije o podatkih. Vztrajna homologija, ki smo jo uporabljali v naši nalogi, uspešno eliminira šum v podatkih in ohranja pomembne informacije iz podatkov.

Implementirana metoda je zelo fleksibilna in jo se da uporabiti na različnih podatkovnih množicah ter v različnih metričnih prostorih.

## 5.1 Bodoče raziskave

Tekom implementiranja metode in analiziranja podatkov o genski izraženosti rakavega tkiva se je odprlo še nekaj novih idej o bodoči raziskavi. V tem delu bomo izpostavili dve najbolj zanimivi, katere bomo v prihodnje lahko realizirali.

Topološke značilnosti dimezije 0 nam povedo število povezanih komponent (skupin), zaradi cilja magistrske naloge, tj. odkrivanja podskupin v posameznem tipu raka, pa je bila naša pozornost usmerjena samo na analizo teh topoloških značilnosti. Ena izmed možnih bodočih analiz bi bila interpretacija topoloških značilnosti dimenzije 1 in 2. Na ta način bi lahko z

nadaljnjo analizo teh topoloških značilnosti prišli do nekaterih pomembnih informacij o bolnikih ter njihovem času preživetja.

Bolj zanimiva možnost bi bila uporaba implementirane metode na genih in ne na bolnikih. Na ta način bi pridobili bolj natančen ,,opis", kako se geni med seboj povezujejo v skupine ter tako poskusili določiti funkcijo genov v skupinah. Pri metodah genskega zdravljenja, igra določanje funkcije gena zelo pomembno vlogo. To pokazuje, da bi bila implementirana metoda v raziskovanju tega tipa uspešna.

# Literatura


[1] N. Bell in A. N. Hirani. PyDEC: Algorithms and software for Discretization of Exterior Calculus. *ACM Transactions on Mathematical Software*, 3:1 – 3:41, 2012.

[2] G. Carlsson. Topology and data. Technical report, Bull. Amer. Math. Soc., 263 – 268, 2008.

[3] N. Chambwe, M. Kormaksson, H. Geng, S. De, F. Michor, N. A. Johnson, R. D. Morin, D. W. Scott, L. A. Godley, R. D. Gascoyne, A. Melnick, F. Campagne, in R. Shaknovich. Variability in dna methylation defines novel epigenetic subgroups of dlbcl associated with different clinical outcomes. *Blood* 123(11):1699–1708, 2014.

[4] D. Cohen-Steiner, H. Edelsbrunner, in J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.* (1):103–120, January 2007.

[5] H. Edelsbrunner in J. L. Harer. *Computational Topology, An Introduction.* American Mathematical Society, ISBN 978-0-8218-4925-5, 2010.

[6] H. Edelsbrunner in D. Morozov. Persistent homology: Theory and practice, 2012.

[7] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, in A. Singh. Confidence sets for persistence diagrams. *Annals of Statistics*, 2014.

[8] J. Friedman in E. J. Alm. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, (9):1–11, 2012.

[9] P. Giblin. Frontmatter. *Graphs, Surfaces and Homology*, Cambridge University Press, tretja izdaja, 1-4, 2010.

[10] T. H.M. Keegan, L. A.G. Ries, R. D. Barr, A. M. Geiger, D. V. Dahlke, B. H. Pollock, W. A. Bleyer, for the National Cancer Institute Next Steps for Adolescent, and Young Adult Oncology Epidemiology Working Group. Comparison of cancer survival trends in the united states of adolescents and young adults with those in children and older adults. *Cancer* (7):1009–1016, 2016.

[11] I. Kosti, N. Jain, D. Aran, A. J. Butte, in M. Sirota. Cross-tissue analysis of gene and protein expression in normal and cancer tissues. *Scientific Reports* (1):6:10 – 6:25, 2016.

[12] E. Laas, P. Mallon, F. P. Duhoux, A. Hamidouche, R. Rouzier, in F. Reyal. Low concordance between gene expression signatures in er positive her2 negative breast carcinoma could impair their clinical application. *PLoS ONE* (2):1–12, 02 2016.

[13] L. Li, W.-Y. Cheng, B. S. Glicksberg, O. Gottesman, R. Tamler, R. Chen, E. P. Bottinger, in J. T. Dudley. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Science Translational Medicine* (311):311ra174–311ra174, 2015.

[14] W. K. Lim, K. Wang, C. Lefebvre, in A. Califano. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics* (13):i282–i288, July 2007.

[15] Y. Liu, Q. Gu, J. P. Hou, J. Han, in J. Ma. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics* (1), 2 2014.

[16] G. Máté in D. W. Heermann. Statistical analysis of protein ensembles. *Frontiers in Physics* (20), 2014.

[17] M. Nicolau, R. Tibshirani, A.-L. Børresen-Dale, in S. S. Jeffrey. Disease-specific genomic analysis: identifying the signature of pathologic biology. *Bioinformatics* (8):957–965, 2007.

[18] L. Seemann, J. Shulman, in G. H. Gunaratne. A robust topology-based algorithm for gene expression profiling. *ISRN Bioinformatics*, 2012.

[19] A. Shahbazi, A. F. Tappenden, in J. Miller. Centroidal voronoi tessellations; a new approach to random testing. *IEEE Trans. Softw. Eng.* (2):163–183, 2013.

[20] C. P. Simmons, F. Koinis, M. T. Fallon, K. C. Fearon, J. Bowden, T.S. Solheim, B. H. Gronberg, D. C. McMillan, I. Gioulbasanis, in B. J. Laird. Prognosis in advanced lung cancer – a prospective study examining key clinicopathological factors. *Lung Cancer* (3):304 – 309, 2015.

[21] A. Singh, A. Yadav, in A. Rana. Article: K-means with three different distance metrics. *International Journal of Computer Applications* (10):13–17, 2013.

[22] L. Song, P. Langfelder, in S. Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* (13):328, 2012.

[23] H. Vikalo, F. Parvaresh, S. Misra, in B. Hassibi. Sparse measurements, compressed sampling, and DNA microarrays. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 581–584, 2008.

[24] H. Wao, R. Mhaskar, A. Kumar, B. Miladinović, and B. Djulbegović. Survival of patients with non-small cell lung cancer without treatment: a systematic review and meta-analysis. *Systematic Reviews* (1):1–11, 2013.

[25] M. D. Wilkerson, X. Yin, K. A. Hoadley, Y. Liu, M. C. Hayward, C. R. Cabanski, K. Muldrew, C. R. Miller, S. H. Randell, M. A. Socinski, A. M. Parsons, W. K. Funkhouser, C. B. Lee, P. J. Roberts, L. Thorne, P. S. Bernard, C. M. Perou, in D. N. Hayes. Lung squamous cell carcinoma mrna expression subtypes are reproducible, clinically important, and correspond to normal cell types. *American Association for Cancer Research* (19):4864–4875, 2010.

[26] W. Zhou in H. Yan. Alpha shape and delaunay triangulation in studies of protein-related interactions. *Briefings in Bioinformatics* (1):54–64, 2014.

[27] A. Zomorodian. Fast construction of the vietoris-rips complex. *Computers and Graphics* Shape Modelling International (SMI) Conference (3):263 – 271, 2010.

[28] A. Zomorodian. *Advances in Applied and Computational Topology.* American Mathematical Society, Boston, MA, USA, 2012.

[29] A. J. Zomorodian, M. J. Ablowitz, S. H. Davis, E. J. Hinch, A. Iserles, J. Ockendon, in P. J. Olver. *Topology for Computing (Cambridge Monographs on Applied and Computational Mathematics).* Cambridge University Press, New York, NY, USA, 2009.